マルチモーダル潜在的ディリクレ配分法の多層化による知識の確率的表現

- ロボットの言語獲得・行動決定への応用 -

ATTAMIMI MUHAMMAD

電気通信大学大学院情報理工学研究科

博士（工学）の学位申請論文

2015年3月

マルチモーダル潜在的ディリクレ配分法の多層化による知識の確率的表現

- ロボットの言語獲得・行動決定への応用 -

博士論文審査委員会

主査　長井 隆行　教授
委員　金子 正秀　教授
委員　田中 一男　教授
委員　横井 浩史　教授
委員　内田 雅文　准教授

Multilayered Multimodal Latent Dirichlet Allocation for Probabilistic Knowledge Representation

Application to Language Acquisition and Action Decision by Robots

Muhammad Attamimi

Abstract

In recent years, studies aiming at the coexistence between robots and humans have been conducted actively. In the field of robot technology, various robots have been developed. However, most are designed to perform particular tasks in limited environments. Furthermore, the actions and responses to inputted patterns required to perform a given task are programmed in advanced by human. To interact naturally with humans, a robot needs to understand human words and act based on the meaning behind those words. Moreover, it is desirable for robots to express their intentions through language generation in communication with humans. To realize these abilities, work has been performed on the symbol grounding problem in the field of intelligent robotics. However, and to the best of our knowledge, a satisfactory solution has still not been found. The goal of this study is to make a breakthrough by considering the understanding and/or generation of language through various concepts formed based on the multimodal information obtained by robots in daily life. Here, "concept" is defined as a "category," such as an object or motion formed using multimodal categorization, whereas various inferences, including the recognition of unseen information, are defined as "understanding." This can be achieved using various formed concepts.

In addition, language can be considered as phoneme labels that connect such concepts. Language acquisition can be achieved through interactions with humans. To this end, a stochastic model is proposed to allow the formation of various concepts through hierarchical multimodal categorization. The proposed method is

based on the graphical model, and categorization can be achieved by estimating the model parameters. The connections between words and concepts are learned using mutual information, whereas grammar can be acquired by considering the order of concepts in human words. Finally, language understanding and/or generation by robots can be realized by incorporating various concepts, connections between concepts and words, and grammar.

Object categorization using multimodal information was proposed by Nakamura *et al.* In fact, the possibility was shown for human-like object concept formation based on information obtained from robot experiences. Robot understanding of an object is possible within the scope of the definition aforementioned based on the prediction of unseen information made through concepts formed by categorizing those experiences. However, category recognition and/or object inference are not sufficient for a robot to act in a human-like intelligent manner. This is because most of the objects are related to the person who uses them, the movement of the objects, and the location where the objects are used. In other words, it is difficult to consider object understanding without regarding prediction on such information. Thus, different types of concepts, such as motion (i.e., movement while using objects), must be captured, as well as the relationship between them. Here, various concept acquisitions can be realized by extending to the hierarchical categorization of multimodal information. Finally, "a computational model of real-world understanding" by the robot can be clarified through this fact. This is the goal of this study.

In chapter 2 of this study, a cleaning task performed by humanoid robots is used as an example, considering that robots are expected to work in domestic environments. To perform the cleaning task, "cleaning" has to be defined. Then, a visual recognition system and planning are implemented to realize the task according to its definition. It is possible to perform the task, including object recognition and grasping actions, within the definition scope. However, cleaning tasks in an unknown environment cannot be realized. This result brings the author to reconsider

the essential meaning of "cleaning." For example "vacuuming," can be thought as of an action to move the vacuum cleaner on small garbage, which is a concept formed from the mutual relationship of "vacuum cleaner" (object concept) and "move something on" (motion concept). Therefore, "cleaning" can be considered as a concept formed from the hierarchical interdependence of various concepts. The formation of these various concepts and the construction of their hierarchical structure are important as the knowledge of robots.

Based on the discussion in chapter 2, we propose, in chapter 3, a probabilistic knowledge representation for robots based on a hierarchical categorization method of multimodal information. The proposed multilayered multimodal latent Dirichlet allocation (mMLDA) is a hierarchy model of multimodal Dirichlet allocation (MLDA). mMLDA consists of bottom layers that include object, motion, place, and person concepts, and a top layer that contains integrated concepts. The following are some examples that use this model at low-level concepts: beverage is an object concept, putting something in the mouth is a motion concept, and dining is a place concept. At the top level, an example of the relationship among these concepts is to learn to form "drinks" as the action concept. From this, an example of inference of unseen information is the inference of "drinks" as an action to be made when observing beverages. Another example is the inference of dining room, which is a place where "drinks" should be performed.

In chapter 4, a method for representing a scene from a sentence is considered using the words and grammar acquired while utilizing various formed concepts. This involves the grounding of word meanings in various concepts with a hierarchical structure, as well as grammar learning. Because this information is not explicitly included in the teaching utterance, a criterion to determine the connection is necessary in the learning algorithm. To this end, the automatic estimation of the connection that uses mutual information between words and concepts is proposed. From this, the connection between words and concepts can be learned to allow inference of the correspondence object, motion, person, and place concepts

of each word. Thus, the order of concepts in teaching utterances can be learned using a simple Markov model that corresponds to grammar. Therefore, language understanding and language generation by robots can be realized.

Furthermore, actual communication is difficult to achieve without considering context, such as background knowledge and surrounding circumstances. In other words, it is necessary to use the various concepts learned and consider the context for more flexible understanding. In chapter 5, a method to determine appropriate actions by integrating various concepts and contexts is proposed. For example, assume that a robot knows that when a person is watching TV on a sofa, he/she usually eats snacks and drinks tea. An appropriate action can be made by the robot using the context "A person is watching TV and drinking tea on the sofa," even if speech recognition errors occur when the person orders "Bring me a snack" to the robot. In chapter 6, the summary and future work of this study is described.

マルチモーダル潜在的ディリクレ配分法の多層化による知識の確率的表現

- ロボットの言語獲得・行動決定への応用 -

ATTAMIMI MUHAMMAD

概要

近年，ロボットと人の共存を目指すための研究が盛んに行われている．現状のロボット技術において，様々なロボットが開発されているが，限られた環境で特定のタスクを実行するものが殆どであり，タスクに必要な行動や入力パターンに対する応答などを人が全て事前に与えなければならない．ロボットが人と自然に暮らすためには，人の言葉を理解する必要があり，その言葉の背後にある潜在的な意味を解釈して行動しなければならない．また，コミュニケーションのために，ロボット自身の意図を言語として創出することが望まれる．旧来の人工知能の研究では，単語を単なる記号として扱い，その記号で閉じた世界の中で言語を理解する努力を続けてきた．自然言語処理・理解は，この流れを強く受けている．これに対して近年のロボティクス・人工知能研究では，いわゆる記号接地問題を基本として，言語の本質的な意味を扱い始めているが，未だに言語の理解や生成の本質的な解決には遠く及ばない．本論文では，ロボットが経験によって得るマルチモーダル情報に基づいて多様な概念を形成し，この概念を基盤とした言語理解・生成を考えることでこの問題を解決する新たな方向性を示す．ここで，概念とはマルチモーダルな情報を分類して形成される「カテゴリ」であり，この概念を通して様々な予測をすることが「理解」であると定義する．さらに言語は，こうした概念と結び付いた音韻ラベルであり，人との自然なインタラクションの中で獲得することが可能である．つまり本論文で提案するモデルは，ロボットが日常の活動によって得ることのできる情報を基盤に概念を形成し，音韻ラベルとの結び付きや語の順番を意味する文法をボトムアップに獲得することで，言語の意味理解や生成を実現するものである．

これまで，マルチモーダル情報を用いた物体のカテゴリ分類手法は中村らによって提案されており，実際に，ロボットが経験することによって得た情報をカテゴリ分類することで，人間の感覚に近い物体概念の形成が可能であることを示している．また，形成された概念を利用して未観測情報を予測することができ，ロボットによる物体の理解が前述の定義の範囲で可能であると言える．しかし，より人間のように柔軟な理解をロボットで実現するためには，物体概念の獲得だけでは不十分であることは明らかである．なぜなら，ほとんどの物体はそれを使う人や使う人の動き，使われる場所などが関連しており，これらの情報を予測できない限りその物体を理解したとは言えないためである．つまり，物体概念のみならず人の動き概念や場所概念など多様な概念を学習すると同時に，それらの関係性を獲得する必要がある．このような多様な概念の獲得は，マルチモーダル情報の階層的カテゴリ分類へと発展させることで実現することで可能であり，最終的にはこれがロボットによる「事物の真の理解の計算モデル」となることを明らかにする．これが本論文のゴールである．

本論文ではまず，第 2 章でロボットが家庭環境で作業することを考慮し，これまで著者が開発したヒューマノイドによる掃除タスクを一例として取り上げる．掃除タスクを行うために,「掃除」を定義する必要があり，その定義に従ったタスクの実現に必要な視覚認識システムやタスクの制御などを実装する．これによって定義範囲内の物体認識や把持行動などを実現することができるが，未知な環境に対して柔軟にタスクを行うことができない．この結果を踏まえて,「掃除」の本質的な意味を考察する．例えば,「掃除機をかける」という行動は掃除機を持って細かいごみの上で動かすことであると考え,「掃除機」という物体概念,「何かの上で動かす」という動き概念の相互関係から形成される概念であると考えることができる．すなわち,「掃除」とは多様な概念の階層的な相互依存関係から構成される概念であると考える．こうした多様な概念の形成とそれらの階層的な構造の構築がロボットの知識として重要である．

第 2 章での議論に基づき第 3 章では，ロボットの確率的知識表現のためのマルチモーダル情報の階層的カテゴリ分類手法を提案する．提案手法は，マルチモーダル潜在的ディリクレ配分法（Multimodal Latent Dirichlet Allocation：MLDA）

を階層化した多層マルチモーダル潜在的ディリクレ配分法（multilayered MLDA：mMLDA）である．下層の MLDA では下位概念である，物体，動き，場所，人物の概念がそれぞれ形成され，上層の MLDA ではこれらの概念を統合する上位概念が形成される．このモデルを用いることで例えば，下位概念としてジュースという物体概念や物を口に運ぶという動き概念，ダイニングという場所概念などが形成される．上位層ではこれらの関係性が学習され,「飲む」という行動概念が形成される．これにより，ジュースを見ることでそれを口に運ぶ「飲む」という行動や，その「飲む」という行動が「ダイニング」という場所で行なわれやすいといった未観測情報の予測を行うことが可能となる．

第 4 章では，形成された多様な概念を利用し，同時に語意や文法を獲得することで，観測したシーンを文章で表現する手法を検討する．ここで扱う問題は，階層的な概念における語意の獲得であり，どの階層のどの概念にどの単語が結び付くかという問題を解く必要がある．本論文では，単語と概念間の相互情報量を用いることで，どの単語が本来どの概念に結び付いているのかを自動的に推定する手法を提案する．これにより単語と概念の結び付きを学習することが可能であり，各単語に対応する，物体，場所や人などといった概念クラスの推定が可能である．従って，教示発話における概念クラスの生起順を学習することで，概念クラスの遷移確率という形で表現される確率文法を学習することができる．これによって，ロボットによる言語の意味理解や生成を実現することが可能となる．

一方，実際のコミュニケーションは，背景知識や周辺の状況などといった文脈を考慮しなければ成立しない．つまり，事物に対する理解をより柔軟に行うためには，学んできた多様な概念を活用した上で，様々な文脈を考慮する必要がある．第 5 章では，ロボットが人と生活する上で，様々な文脈においてどのように行動決定するかを議論する．つまり，獲得した多様な概念と文脈と統合することで，適切な行動を決定する手法を提案する．これにより例えば，人が普段ソファーでテレビを見ているときに，お菓子を食べながらお茶を飲んでいるということを知っていれば，人が「お菓子を持ってきて」と命令した際の音声認識に誤りが生じたとしても，そのときに「ソファーでテレビを見ていてお茶を飲んでいる」という文脈を用いることで，ロボットが適切に判断をして正しい行動をとることができ

る可能性がある．第 6 章では，本論文のまとめと今後の課題について述べる．

目 次

図 目 次

表 目 次

略語一覧

DPM	Dirichlet Process Mixture
DSIFT	Dense Scale Invariant Feature Transform
EM	Expectation Maximization
GMM	Gaussian Mixture Model
GPSR	General Purpose Service Robot
HDP	Hierarchical Dirichlet Processes
HDP-HMM	Hierarchical Dirichlet Processes Hidden Markov Model
HMM	Hidden Markov Model
ICP	Iterative Closest Point
LDA	Latent Dirichlet Allocation
LRF	Laser Range Finder
MFCC	Mel-Frequency Cepstrum Coefficient
MHDP	Multimodal Hierarchical Dirichlet Processes
MHDP-HMM	Multimodal Hierarchical Dirichlet Processes Hidden Markov Model
MLDA	Multimodal Latent Dirichlet Allocation
pLSA	probabilistic Latent Semantic Analysis
RRT	Rapidly Exploring Random Tree
SD	Shape Distribution
SVM	Support Vector Machine
TOF	Time Of Flight

第1章　序論

1.1　はじめに

本来ロボットの役割は，人間の代わりに何らかの作業を行うことである．この役割が追及された究極の形が産業用ロボットであろう．こうしたロボットは，ティーチングプレイバック方式で動作するのが通常であり，全ての動作や入力パターンに対する応答が全て事前に準備される必要がある．近年ではロボットが我々の身の回りに進出しており，その代表格が掃除用ロボットである．掃除用ロボットは，業務用のものもあれば，Roomba [1] に代表される家庭用も存在する．しかし依然としてこれらのロボットは，制限された環境において決められた掃除タスクを上手にこなすことができるものの，それ以外の環境やタスクなどに対応することは困難である．

一方で，人や動物に近い形や機能を持つロボットも盛んに研究開発されている．例えば，SONY の AIBO（Artificial Intelligence Robot）[2] が代表的なペットロボットであり，今でも一部には根強い人気があると言われている．しかし当時，発売後の数週間で多くのユーザが飽きてしまったという報告もある．これは，犬のような賢い振る舞いが実は決められた行動であって，多くのユーザがこれに気付いてしまったためであると考えられる．またパロ [3] のようなセラピーロボットも注目されているが，基本的には決められた振る舞いしか行うことができない．ヒューマノイドロボットとしては，HONDA の ASIMO（Advanced Step in Innovative Mobility）[4] や TOYOTA の TPR（Toyota Partner Robots）[5] が有名である．ASIMO は，走ったり，階段を登ったり，踊ったりすることができる．また TPR は，トランペットを吹いたり，ドラムを叩いたりすることができる．さらにここ数年は，災害現場で活躍するヒューマノイドロボットが注目されており，DARPA（Defense Advanced

Research Projects Agency）ロボティクスチャレンジ [6] での，ボストンダイナミクス社 Atlas の活躍は記憶に新しい．しかしながら，現状 DARPA ロボティクスチャレンジで活躍しているロボットは遠隔操作されており，そうした未知の環境で自律的にタスクをこなすことは困難である．このような非常に高度で複雑な身体をもったロボットでさえ，人間のように柔軟に物事に対処することは非常に難しいのが現状である．

1.1.1 RoboCup@Home におけるタスク

「ロボットが未知の環境で，どれだけ柔軟に我々の身近なタスクをこなせるか?」という視点での取り組みとしては，RoboCup@Home [7] が存在する．Robocup@Home は特に家庭内タスクをテーマとし，会場に設営された未知の家庭環境でロボットが家庭内タスクを実行し，その成功度を点数として競うプロジェクトである．そこでは，人から飲み物の注文を受けてその人に届けるというタスクや，部屋内に雑多に置かれた物を片付けるタスクなどがある．RoboCup@Home における現状のマイルストーン的課題は，GPSR（General Purpose Service Robot）と呼ばれるタスクである．これは，コンピュータがランダムに生成した命令文を人がロボットに命令し，その命令がどれだけロボットによって実行されるかを競う．このタスクには言語理解が必要であり，当然のことながらこれを完璧にこなすことのできるロボットは今のところ存在しない．

実際著者は，Robocup@Home に参加し競技のためのタスクの開発に携わった．開発したタスクの一例として，ヒューマノイドロボットによる掃除タスクがある．掃除タスクを行うためにまず,「掃除」を定義する必要があった．そこで著者は掃除タスクとは,「掃除の対象となる卓上を，ロボットが初めて記憶した状態に戻すこと」と定義することとした．この定義に従ってタスクを達成するため，物体認識システムを開発しプランニングに必要な行動を予め用意する必要があった．このタスクでは，卓上という限定された環境において定義範囲内の行動を実行することで掃除タスクを行うことができる．しかし，明らかにこれは作り込みであり,「柔軟にタスクをこなせるか？」という問いに答えてはいない．

以上のように，これまで開発されてきたロボットはある決められた環境で事前

にプログラムされたタスクを実行することには長けている．こうした決められたことを正確に行うための仕組みは，データベースに記述されたルールとのマッチングに基づく認識である．しかしこれは，人間が日々行っている柔軟な認識（知能）からは程遠いと感じる．人間のような柔軟に考え行動するロボットを実現するために，このような認識システムは不十分であり，人間のように柔軟に物事を理解するシステムが必要であると考える．そこで本論文では，事物を真の意味で理解することのできるロボットの実現について議論する．そのためにロボットは概念を獲得し，その概念に基づいて様々なことを認識・理解する仕組みが必要なのではないかと考える．本章ではまず，人間における理解から始め，その議論に基づきロボットの知能の構成を試みる．

1.1.2 人間における理解

「人間がどのように物事を理解しているのか?」という問いに対して，人間の知性という文脈で古くから研究がなされている．例えば，1689 年に出版された「人間悟性論」[8] において Locke は，人間の知性がどのような対象を扱うのに適しているのか，またはいないのかを明らかにしようとした．Locke の主張の一つは，観念が発生する以前の心の状態は白紙であり，これをタブラ・ラサと呼ぶ．タブラ・ラサの考え方は，生得的な能力を否定するものであり，観念自体が複雑であっても全て経験に由来するものとして考えることになる．つまり，外界より入力された知覚情報とそれへの心理的作用により観念が発生しており，それが知性に貢献することになる．現代の認知・発達心理学の知見は，タブラ・ラサの考え方に否定的であるが，生後の経験によってボトムアップに知性が発達していくという様相は，本論文の基本的な姿勢に通ずるものである．また Locke は，観念には単純観念とそれらを組合せた複合観念があるとしている．さらには観念の記憶として言語があり，言語を観念の典型的または抽象的な形態として使用させる手段と考える．本論文では，概念とその複雑な関係性を表現する概念構造が知性にとって重要であるという基本的な理念に基づいて議論を進めて行くが，これはまさに Locke の思想に関連している.「人間が思考するとき、知性の対象となるもの」を Locke は「観念（idea)」と呼び，これが本論文で定義する概念と全く等価であるわけで

はないが，これらが類似していることは明らかであろう．また，単純観念とそれらを組み合わせた複合観念の重要性は，本論文のこれからの議論に重要な示唆を与えている．

文献 [9] によると，概念は人の精神生活やコミュニケーションの基本となる．多くの認知科学者は概念がメンタルな表現であり，全体から選択されるカテゴリとなっていることを認めている．また文献 [10] では，日常生活において人間のカテゴリ分類が重要な役割を果たすことが報告されている．人間はカテゴリを形成することで，経験した物事を全て参照することなく，必要最小限の認知的処理によってより多くの情報を得ることができる [11]．このカテゴリには，単純なカテゴリやそれらを組合せた複雑なカテゴリが存在し，それらは階層的な構造を持つ．階層的に分類されたカテゴリを利用することで，未知な事物に対する推論が可能となり，行動決定や問題解決に大きく貢献する [12, 13]．さらに，人間の言語獲得を説明する理論として,「制約論」[14] が Markman によって提案される．この理論の重要な点は，概念化された事物に言語（単語）を結び付けることで，言語獲得の仕組みが説明できることである．また，獲得した言語と多様な概念を利用することで，人はある事物を言葉として表現することができる．

1.1.3 人工知能における理解

人工知能の分野においては長い間，人間の知能を計算機が扱うことのできる記号の操作としてモデル化してきた．このモデルでは，入力する信号やそれを処理する能力は全て記号で表現され，数学的・論理的に記述される．このような人工知能は，Symbolic AI（Artificial Intelligent）とも呼ばれ，成功した一例として有名なのは，IBM の Deep Blue である．Deep Blue は，チェスの世界チャンピオン Kasparov に勝つことができた．チェスのようなゲームの世界は記号表現が可能であるため，計算機の能力が人間を凌駕する一方で，実世界の人間の活動の場は必ずしも記号で表現できる訳ではなく，こうした人工知能が人間以上に機能するのは現状では難しい．これは人工知能における基本問題であり，記号接地問題と呼ばれる．

記号接地問題は Harnard が取り上げた問題であり,「記号がどのように実世界と

結び付いているか?」を問う問題である．Harnard は，ある系が操作する記号が単に形式的なものではなく意味のある記号であるためには，記号の対象が存在している実世界に，記号が接地されている必要があると主張した [15]．しかしこの問題に対して，Harnard は具体的な解決方法を提案したわけではない．一方，記号を意味付けするアプローチとしては，オントロジーや意味ネットワークなどがある [16,17]．これらのモデルは，記号（単語）と別の記号の関係を表現するモデルとなっている．つまり言い換えればこれらのモデルは，言葉（単語）を別の言葉で説明するものとなる．そのため本質的にはこれらの手法も，記号接地問題の解決になっていない．また，オントロジーや意味ネットワークは人が設計する必要があるという点において大きな問題を抱えることになる．

この問題をどのようにすれば解決できるかについて，上述の人間における理解という観点から考えてみたい．例えば,「りんご」という言葉の意味を理解するために，人は実世界に存在する対象としての「りんご」を自分の目で見たり手で触ったり，時には嗅覚や味覚を使って得ることのできる知覚情報から「りんご」という概念を形成する．そしてこの概念と音韻列である「りんご」とを結び付けることで初めて,「りんご」という言葉の意味を理解することができる．このように本論文では，実世界に存在する対象を感覚情報を通して概念化し，形成された概念と音韻列を結び付けることで記号接地問題を解決することを考える．これはつまり，ロボットが自身の身体性によって得る経験を通して概念を形成し，その概念と言語を結び付けることで言語の意味を理解する仕組みである．また，こうした概念構造はあくまでもボトムアップに教師なしで形成されるものであり，設計者が設計するのではないことが重要である．

1.1.4 マルチモーダル情報の階層的カテゴリ分類による事物の理解

これまでの先行研究で，上述のような議論に基づき，マルチモーダル情報を用いた物体のカテゴリ分類手法が提案されている [18]．マルチモーダル情報は，ロボットが自身の身体を用いて実際に物体を見て，掴んで振ることで得られる視覚，触覚及び聴覚情報である．これらの情報をロボットが自律的にカテゴリ分類することで，人間の感覚に近い物体カテゴリを生成できることが示されている．ここ

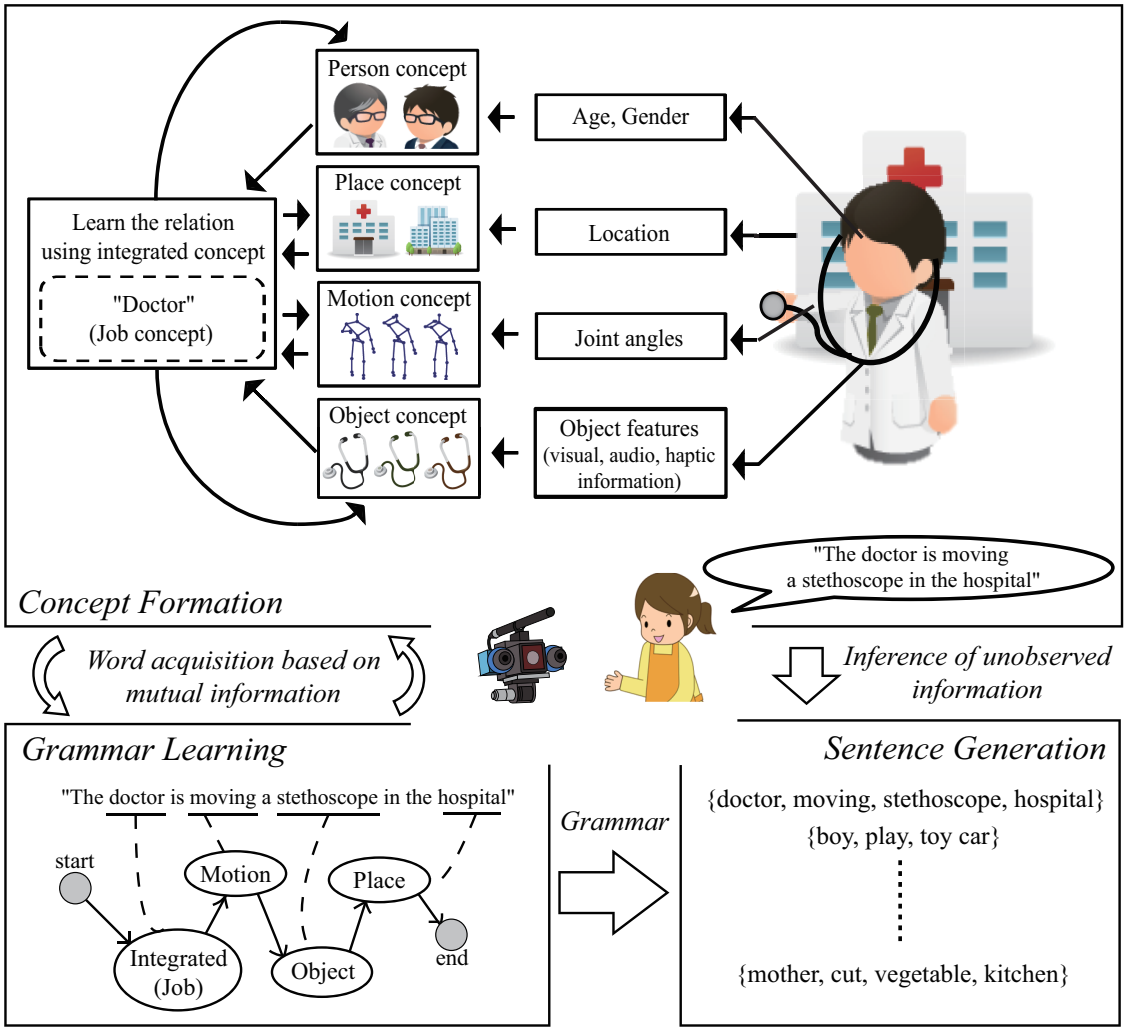

図 1.1: 多様な概念の獲得と言語理解

での重要なポイントは，学習された物体カテゴリをベースとした未観測情報の予測であり，これがロボットによる理解につながることである．本論文では，ロボットが教師なしでマルチモーダル情報をカテゴリ分類することによって形成されるカテゴリを概念と呼ぶ．また，先行研究で形成される物体に関するカテゴリを「物体概念」と呼ぶ．

しかし，より人間の様な理解をロボットで実現するためには，物体概念の獲得だけでは不十分であることは明らかである．なぜなら，ほとんどの物体は，それ

を使う人や使う人の動き，使われる場所などが関連しており，これらの情報を予測できない限り，その物体を理解したとは言えないためである．つまり，物体概念のみならず人の動き概念や場所概念など多様な概念を学習すると同時に，それらの関係性を獲得する必要がある．このような多様な概念の獲得は，先行研究の知見である単一概念の形成を基盤として，マルチモーダル情報の階層的カテゴリ分類へと発展させることで実現する．最終的には，これがロボットによる事物の真の理解の計算モデルとなることを明らかにしたい．

本論文では，多様な概念をそれぞれ獲得すると同時に，それらの関係性を表すより高いレベルの概念を形成することを考える．本論文で中心となるアイディアを模式図にしたものが，図 1.1 である．この図では，日常生活を経験してきたロボットが,「40 代の男性（人物概念）」,「何かを動かす（動き概念）」,「聴診器（物体概念）」,「病院（場所概念）」という 4 つの下位概念を使って発話を理解する様子を示している．さらにこれらの概念が統合されることで，より高いレベルの概念である「医者」（統合概念で，この場合「仕事」を意味する）という概念が形成される．この図において重要なのは，様々なレベルでの推論（予測）が可能であるということである．上記の例においてロボットは，与えられた「聴診器」の視覚的な情報から,「何かを動かす」という動きや「病院」という場所などを想起することができる．勿論，動きや場所など別の情報からの推論も可能である．また，上位概念の形成過程が下位概念の形成，つまりは物体や動きのカテゴリ分類に影響を及ぼすことは注目に値するであろう．例えば，テクスチャが似たような「聴診器」の形をしている物体（ロープなど）は,「聴診器」と同じカテゴリに分類される可能性があるが，この物体が「病院」という場所で使用される場合，統合概念である「医者」が下位層の分類に影響することで,「聴診器」（物体概念）といった単一の物体概念を形成することに寄与する．一方で，物体が異なる場所や動きなどに関係する場合，見た目の異なる物体であっても同じカテゴリに分類される可能性がある．

さらに本論文では，マルチモーダル情報の分類により形成された階層的概念を用いて語意の獲得手法を提案する．先行研究において，入力されたマルチモーダル情報に対応する単語や，単語が指す概念の推論の可能性が示されている．しか

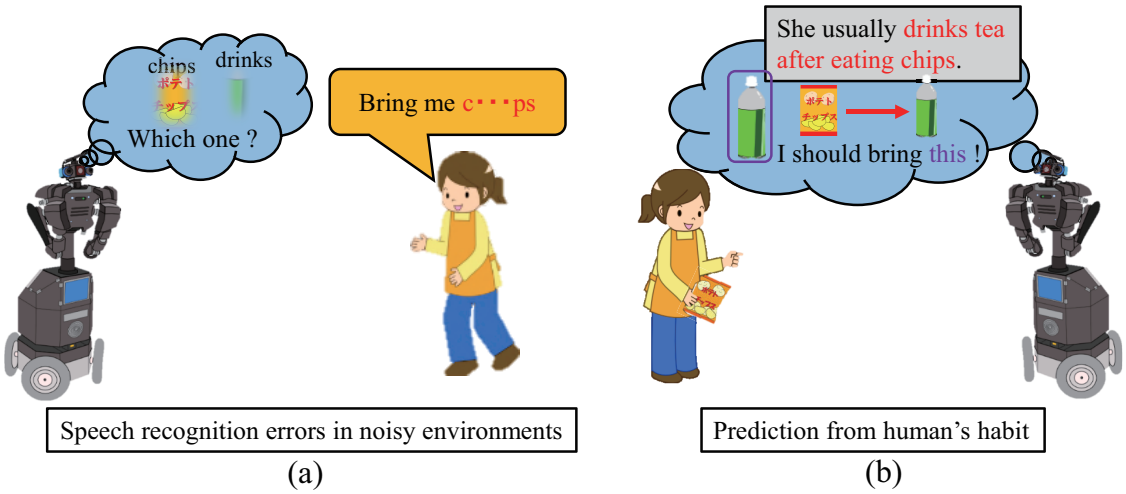

図 1.2: 文脈を考慮したロボットの行動決定

しここで扱っている問題は，階層的な概念における語意の獲得であり，どの階層のどの概念にどの単語が結び付くかという問題を解く必要があり，この問題は，先行研究では実現されていない．本論文では，単語と概念間の相互情報量を用いることで，どの単語が本来どの概念に結び付いているのかを自動的に推定する手法を提案する．これによって，図 1.1 に示すように,「医者は病院で聴診器をあてる」という教示発話より,「医者」（統合概念）,「聴診器」（物体概念）,「病院」（場所概念）,「あてる」（動き概念），といった対応関係が学習される．さらには，教示発話における概念の生起順を学習することで，概念の遷移確率という形で表現される確率文法を学習することができる．本論文では，このように獲得した文法と概念に結び付いた単語を用いて，ロボットが観測したマルチモーダル情報から文章を生成することを考える．

1.1.5 文脈の統合によるロボットの行動決定手法

一方，実際のコミュニケーションでは，背景知識や周辺の状況などといった文脈を考慮しなければ成立しない．つまり，事物に対する理解をより柔軟に行うためには，学んできた多様な概念を活用した上で，様々な文脈を考慮する必要がある．

そこで本論文ではさらに，様々な文脈を統合することで，適切な行動を決定する手法を提案する．一般に，ロボットは人間の命令に応じて行動する．人間の命令は一つの行動に対して様々であり，適切な行動を行うためには，適切に命令を解釈しなければならない．また音声命令では，音声認識誤りが生じる可能性を考慮する必要がある（図 1.2（a））．ロボットが命令を正しく解釈するための手がかりとして，命令を受けた際の文脈が考えられる．例えば，人が普段ソファーでテレビを見ているときに，お菓子を食べながらお茶を飲んでいるということを知っていれば，人が「お菓子を持ってきて」と命令した際の音声認識に誤りが生じたとしても，そのときに「ソファーでテレビを見ていてお茶を飲んでいる」という文脈を用いることで，ロボットが適切に判断をして正しい行動をとることができる可能性がある．また，人が日々行っている行動をロボットが学習できれば，人からの命令がなくても，人の次の行動を予測し，適切なサービスの提供が実現できると考えられる（図 1.2（b））．

1.1.6 関連研究

関連研究としては，センサ情報に基づいた物体のカテゴリ分類に関する研究が挙げられる [19–24]．また，人間の動きのモデル化についても多くの研究がなされている [25–27]．本論文では，知覚情報の分類が主眼であり，その点においては上記の研究とは同様の方向性であると言える．しかし，本論文で提案するモデルでは複数の概念（特に動き概念と物体概念）とそれらの関係性を同時に学習することを目的としているという点で，上記の研究とは大きく異なる．従って提案モデルでは，概念間の推論が可能であるのに対し，これらの研究ではそうした点については考慮されていない．

一方で尾形らは，Parametric Bias を用いた Recurrent Neural Network（RNNPB）を用いることで，異なるモダリティ間の情報をマッピングすることのできる手法を提案している [28]．このシステムは，物体の動きによって生成される音を表現する運動をロボットが生成できるように学習することが可能である．従って論文の目的は，ロボットが異なる種類のセンサからの信号間のマッピングを学習するモデルを，RNNPB によって構築することである．このモデルを用いることで，ロ

ボットは音から関連する動作を生成することができるようになり，これは本論文の目的と非常に関連している．しかしながら，[28] ではカテゴリ（概念）と，それらの相互依存関係を明示的に扱っているわけではなく，複数の概念を統合することにより獲得される上位の概念といったことも考慮していない点で，本論文の提案するモデルとは大きく異なると言える．また RNNPB は，スケーラビリティに問題がある可能性がある．実際，文献 [28] では，5 つの物体で実験を行っているのみであり，物体数などが大幅に増えた際にモデルが実際に機能するかどうかは必ずしも明らかではない．

さらに文献 [29,30] では，感覚運動マッピングとしてのアフォーダンス学習を提案している．この論文では，ベイジアンネットワークを用いて物体・動作・効果の関係性をモデル化している．しかし，提案されているモデルの構造は非常にシンプルであるため，本論文で扱う複雑な概念構造を表現するのは困難である．また，扱う動作は固定されており，ロボットが新規な動きの概念を学ぶことができないという問題もある．これは，与えられた概念間の関係性のみを学習していることに相当していると言える．これに対して本論文では，センサ入力から動きや物体の概念を獲得すると同時に，それら概念間の関係性も同時に学習する枠組みとなっている．

コンピュータビジョンの分野では，human-object interaction（HOI）なる考え方が提案されている [31,32]．これは，人間の動作の認識には使用されている物体が何かということが手掛かりとなると同時に，物体を認識する際に人間の動作や姿勢が重要になるという考え方である．つまり，HOI をモデルに組み込むことで，物体検出および人の姿勢推定の性能を大幅に向上させることができる．しかしこれらの研究は，教師あり学習であり，本論文で扱う教師なしの学習問題とは大きく方向性が異なる．

また，シーンからの文生成は，文献 [26,33] で提案されている．人の運動のモデル化と，運動からの文生成が文献 [26] で提案されているが，物体と動きなどの関係性は考えられていない．文献 [33] では，コーパスを用いた視覚情報からの文生成の枠組みが提案されているが，文の構造が単純である．また，動画中の人の動作を説明する文を生成する研究 [34] や，動画に映る調理の動作を説明する文を生

成する研究 [35] なども存在する．しかしこれらの研究では，視覚情報のみから文生成を行っており，マルチモーダル情報を扱っていない．これらの研究に対して本論文では，多様な概念の形成と統合，言語の学習及び構造が単純でない文生成を扱い，こうしたことがロボットによって教師なしで学習されるという点で大きく異なると言える．

1.1.7 本論文の構成

本論文の構成は以下の通りである．

2 章： ロボットが人と共存するにあたって，人の言葉を理解しその言葉の背後にある潜在的な意味を解釈して行動する必要がある．ロボットの行動（タスク）は単純なものから複雑なものまで様々であり，これらをこなすための知能（知識）が必要であると考える．2 章ではロボットが家庭環境で活動することを前提として，掃除タスクを例に取り上げて概念や言語理解の必要性にについて議論する．実際著者は，RoboCup@Home に参加し競技におけるタスクを実現するために，ヒューマノイドロボットによる掃除タスクを実装した．タスク実装のためにまず掃除タスクを定義し，その定義に従った掃除タスクを実現するためのハードウェア，視覚認識システムやプランニングなどについて述べる．さらに掃除タスクの実装によって，実現できたことや実現できなかったことを通して，ロボットのタスクと概念及び言語理解との関係性について議論する．

3 章： 2 章においてロボットがタスクを行うためには，知識が必要でありその知識を自律的に獲得することが重要であることを議論した．3 章ではまず，ロボットのタスクに最も必要とされる知識である動作概念の獲得手法について述べる．提案手法は，先行研究より提案されたマルチモーダル潜在的ディリクレ配分法（Multimodal Latent Dirichlet Allocation：MLDA）を多層化した多層 MLDA（multilayered Multimodal Latent Dirichlet Allocation : mMLDA）である．mMLDA によって，ロボットのタスクに必要な知識である物体や動き概念と，それらの関係を表現する動作概念の獲得を同時に行うことが可能

であることを示す．さらに，獲得した概念を利用することで，未観測情報や概念間など様々な予測が可能であることを実験通して示す．

4 章： 3 章において提案したモデルでは，物体と動きに対する理解を実現することができるが，より柔軟に理解を行うためには物体や動きの概念だけでは不十分であり，例えば場所や人物など多様な概念を獲得する必要がある．また，獲得した多様な概念の意味をどのように言葉として表現するのかを考える必要がある．4 章では，このような問題を解決するために，mMLDA を物体，動き，場所，人物概念とそれらの関係性を表現する統合概念へ拡張し，獲得した階層的な概念に単語の結び付けを自動的に行う手法を提案する．さらに獲得した語意と教示発話を用いることで，教示文に含まれる概念クラスの生起順を計算し，文法の学習を行う．最終的には多様な概念，語意及び文法を用いて，観測したシーンを文章として表現する手法を提案し，実験を通して提案手法の有効性を示す．

5 章： 4 章において，多様な概念を獲得可能なモデルについて議論し，これが事物の理解において重要な役割を果たすことを明らかにした．しかし，ロボットが人を相手にしてタスク（サービス）を行うために，背景知識や周辺の状況などといった文脈を考慮しなければ，これを適切に行うことができない．5 章では，ロボットが人と生活する上で，様々な文脈においてどのように行動決定するかを議論する．これは mMLDA のさらなる応用として，人の言語命令のみならず，人の習慣（現在の動作と次の動作と動作する際に関係する物体や場所など）を手がかりにサービスするロボットの実現を目指したものである．5 章では，シミュレーション実験を通して提案手法の有効性を示す．

6 章： 本論文のまとめと今後の課題について述べる．

第2章　ロボットのタスクと概念・言語理解

2.1　はじめに

ロボットが人と自然に暮らすためには，人の言葉を理解する必要があり，その言葉の意味を解釈して行動しなければならない．前章において述べた，人のように理解をするロボットは，人と活動する際に得た経験をもとに多様な概念を獲得し，獲得した概念を利用することで，事物に対する様々な予測を行う．また，概念と言葉を結び付けることで，言語の理解が可能となる．

一方で，近年ロボット技術が進歩しており，ロボットが様々な場面・用途に利用されるようになってきた．中でも我々の身近に存在するロボットとして，掃除ロボットの普及が目覚ましい．特に iRobot 社の Roomba [1] が，最も成功していると言える．こうしたロボットの知能とはどのようなものであろうか？明らかに Roomba は，概念や言語理解の仕組みを持っていない．そのようなロボットが大きな成功を収めているということは，概念や言語理解がロボットには求められていないことを意味しているのであろうか？

実際 Roomba は，Brooks の Subsumption Architecture（SA）[36] をベースに作られている．SA 理論では，複雑な知的振る舞い（システム）を多数の単純なシステムに分割し，それらの階層構造を構築する．各層（システム）は，それぞれ独自の入出力が存在し，他の層とは独立に環境と相互作用ができるように作られている．また，システムを構成する各層において，低次の層は高次の層に依存せずに並列的に作動できる．これにより，環境の状況に応じた行動が実現される．Brooks の考え方では，最初にあるのは行動であって，この行動と外界との相互作用から知性が生まれる．単純なタスクにおいて，SA 理論で構築されたロボットは環境の

変化に対してロバストにタスクを行うことができるが，人間が日々行っているような複雑なタスクにおいては，このような知能だけでは不十分であろう．特に，人の言葉（命令）を解釈し動作するようなタスクにおいては，人間の言葉を理解しなければ実現不可能である．逆に Roomba は，非常に単純なタスクをロバストに実行することを目的に設計されており，そのロバストさこそが一般に受け入れられている理由であると言える．

しかし，ロボットに求められるのは単純なタスクをこなすことだけではない．例えば，家庭において求められるタスクに競技形式で取り組む RoboCup@Home [7] では，GPSR（General Purpose Service Robot）と呼ばれるタスクが存在する．GPSR はランダムに生成された言語命令に対して，ロボットがどれだけ適切に行動できるかを競う．このタスクでは，ロボットによる命令の解釈（言語理解）が必要であり，未だこれを完璧にこなすロボットは存在しない．また掃除タスクだけを考えてみても，実際には難しい問題が多く存在する．

そこで本章では，ロボットが家庭環境で動作することを前提として，掃除タスクを例として概念や言語理解の必要性に関する議論を進める．実際著者は，RoboCup@Home におけるタスクを実現するために，ヒューマノイドロボットによる掃除タスクを実装した．ここではまず実装のために掃除タスクを定義し，その定義に従った掃除タスク実現のためのハードウェア，画像認識技術，タスク制御について述べる．そして，その実装によって実現できたことや実現できなかったことを通して，タスクと概念・言語理解との関係性について考察する．

2.2 掃除タスクの概要

ここでは，著者が RoboCup@Home のタスクとして掃除ロボットを実装した例について述べる．タスクを実装するためには，その定義が必要であり，著者は次のように「掃除」を定義することから始めた．本章における「掃除」とは,「対象となる机の上を，きれいな状態（ロボットが最初に見た状態）に戻すことである」と定義する．ここでは対象となる卓上に存在する物体に対して，ロボットが知らなければ（未知物体）「ごみ」と定義する．一方，既知物体はごみでないとし，こ

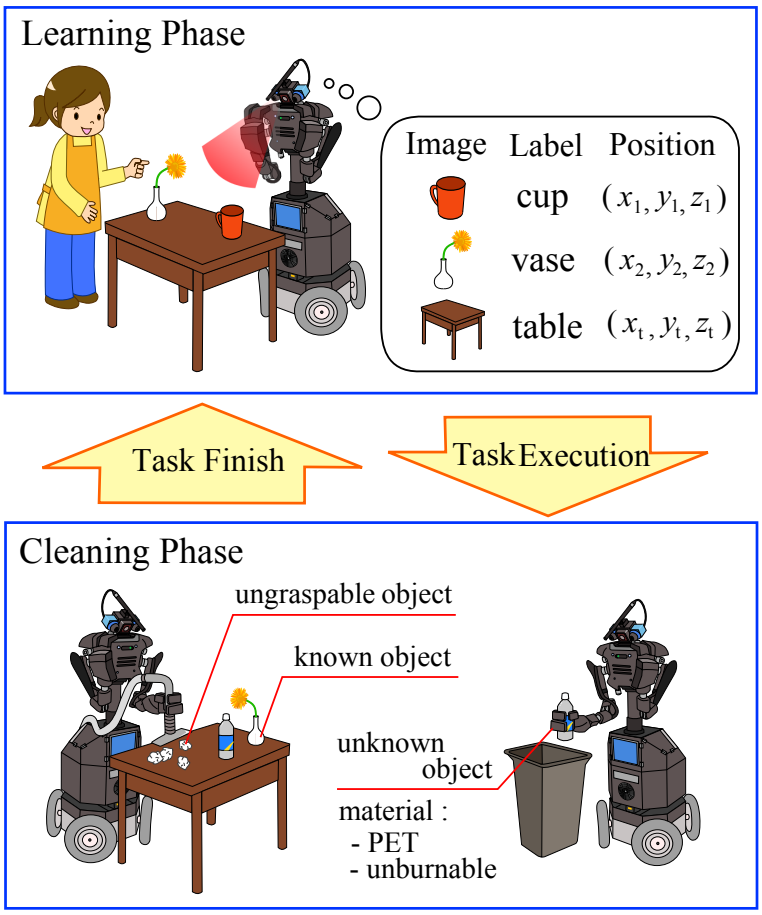

図 2.1: 掃除タスクの概要図．上部は学習フェーズを示しており，掃除タスクの対象である卓上のきれいな状態を記憶する．下部は掃除タスクのメインであり，卓上の認識と物体操作（物体把持及び掃除機による細かい物体の掃除）を含む

こではこれを「商品」と呼び，掃除する際に元の位置に戻す必要があるとする．タスク完了時に，ごみとなる物体はごみ箱に捨てられ，商品となる物体は卓上のもとの位置に置かれていることになる．

卓上の物体は，ロボットの把持能力に応じて，把持可能な物体と把持不可能な物体に区別することができる．把持不可能な物体はさらに，細かい物体と大きい物体に分けることができる．細かい物体とは，高さの低い物体であり，平面検出に基づく手法だけでは検出が困難である．通常この物体は，砂糖や粉などのよう

に軽くて細かいため，机の上の細かい物体のほとんどはごみとして考えることができる．また，細かい物体は軽いため掃除機で吸い取ることが可能である．一方で，電気ポットやコーヒーメーカーなど大きい物体は，平面検出で位置特定が可能であるが，重いため把持できない．通常，大きい物体は商品として分類される．把持可能な物体に対して，把持可能なごみはごみ箱に捨て，把持可能な商品は適切な位置に置く．

掃除タスクの概要を，図 2.1 に示す．図 2.1 中の掃除タスクには，学習フェーズとタスク実行フェーズが含まれる．学習フェーズでは，机の上のきれいな状態を学習する．まず卓上の色と材質を学習し，机の上にあるべき商品を認識し，その位置を記憶する．タスク実行フェーズには，卓上の物体検出と物体認識が含まれている．把持可能な物体は適切な位置に運び，把持不可能な物体を掃除機で吸う．

掃除タスクを行うための要素技術として，ロボットの移動，物体操作と認識が必要である．ロボットの移動は，レーザーレンジファインダ（LRF：Laser Range Finder）を用いた自己位置推定と地図作成（SLAM：Simultaneous Localization and Mapping）により行う．視覚センサ [37] より取得した 3 次元情報を用いて，Rapidly-exploring Random Tree（RRT）[38] に基づいたパスプランニングで物体を操作する．ロボットによる認識は，視覚センサより取得できる色，テクスチャ，3 次元情報と近赤外線反射強度を利用した視覚認識システムを構築することで実現する．視覚認識システムは物体検出，物体認識，材質認識からなり，詳細は次節で述べる．

2.3 ロボットプラットフォームと視覚処理システム

本節では，タスク実現に用いたロボットプラットフォームと視覚処理システムについて述べる．これは，後の節でも用いられるシステムであり今後の議論の前提となるためである．特に視覚情報としてどのようなものが得られるかは，タスクを実現する上でも非常に重要である．

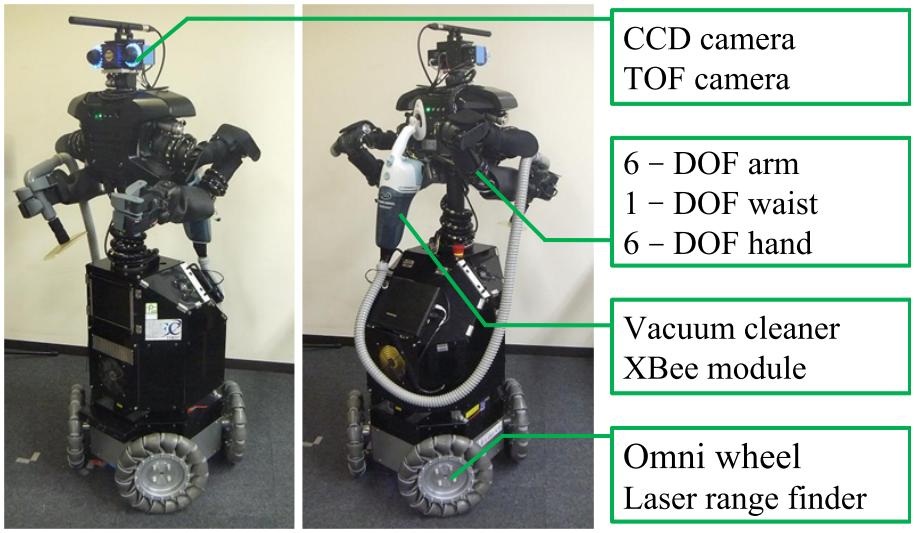

図 2.2: ロボット，視覚センサ及びハンディ掃除機

2.3.1 ロボットプラットフォーム

ここでは，図 2.2 に示すロボットプラットフォームを利用する．このロボットは全方位台車をベースとしており，LRF を用いた自己位置推定と地図作成を行うことができる．自律移動には，RRT [38] に基づいたパスプランニングが用いられる．ロボットの左右にそれぞれ 6 自由度アーム，6 自由度のハンドが含まれる．把持の際のプランニングに関しては，移動のパスプランニングと同様な手法を用いており，これによってロボットが衝突を避け，物体を把持することが可能となる．さらにハンディ掃除機が設置され，そのホースを手に持たせることで卓上のごみを掃除する．この際，片方の手は物を掴むことができるため，缶などのごみは把持してごみ箱に捨てることができる．また，掃除機には XBee 無線モジュールが取り付けられており，ロボットが自ら電源をオン/オフすることができる．頭部には，2 自由度のパンチルト台の上に視覚センサ（詳細は 2.3.3 を参照されたい）が搭載されている．

2.3.2 視覚認識システム

掃除タスクを実現するに当たって必要となる視覚処理は,(1) 物体検出,(2) 物体認識,(3) 材質認識，である．一般に，複雑な環境における (1) は簡単なタスクではない．本論文では (1) において,(a) 動きアテンションによる物体検出,(b) 平面検出による物体検出,(c) アクティブ探索による物体検出 [39],(d) 把持できないごみの検出，を用いる.(a) は複雑なシーンにおける新規物体の学習に用いることが可能であるが，机の上の物体に対しては (b) と (c) が望ましい．また,(b) と (c) を相補的に利用することでよりロバストな物体検出システムが実現できる．

2.3.3 視覚センサ

本章では，文献 [37] で提案された 3 次元センサを用いる．3 次元センサは図 2.2 に示すように，赤外線 TOF（Time Of Flight）カメラと 2 台の CCD カメラから構成される．TOF カメラと CCD カメラのキャリブレーションを行うことで，高速かつ高精度に色情報と 3 次元情報を取得することができる．さらに，TOF カメラより距離情報の信頼度を測定するために使用される近赤外線反射強度を取得することも可能である．従って，3 次元センサより色，テクスチャ，3 次元情報，近赤外線反射強度が得られる．これらの情報を用いて，物体認識，材質認識，物体検出システムを構築する．

2.3.4 複数特徴量を用いた 3 次元物体認識

本節で用いる物体認識手法の概要図を図 2.3 に示す．物体認識において，物体学習フェーズと物体認識フェーズに分けることができる．学習フェーズでは，対象物体を複数の視点から観測し，各観測データに対してそれぞれの特徴量を計算しそれらをデータベースとして保持する．認識フェーズでは，同様に物体を抽出し，その領域内の特徴量とデータベースを比較することで物体を認識する．この際重要なのは，複数の特徴量をどのように統合するかという問題である．本節では，環境の変化やデータベース内で類似した特徴を持つ物体の有無に応じて，自動的に

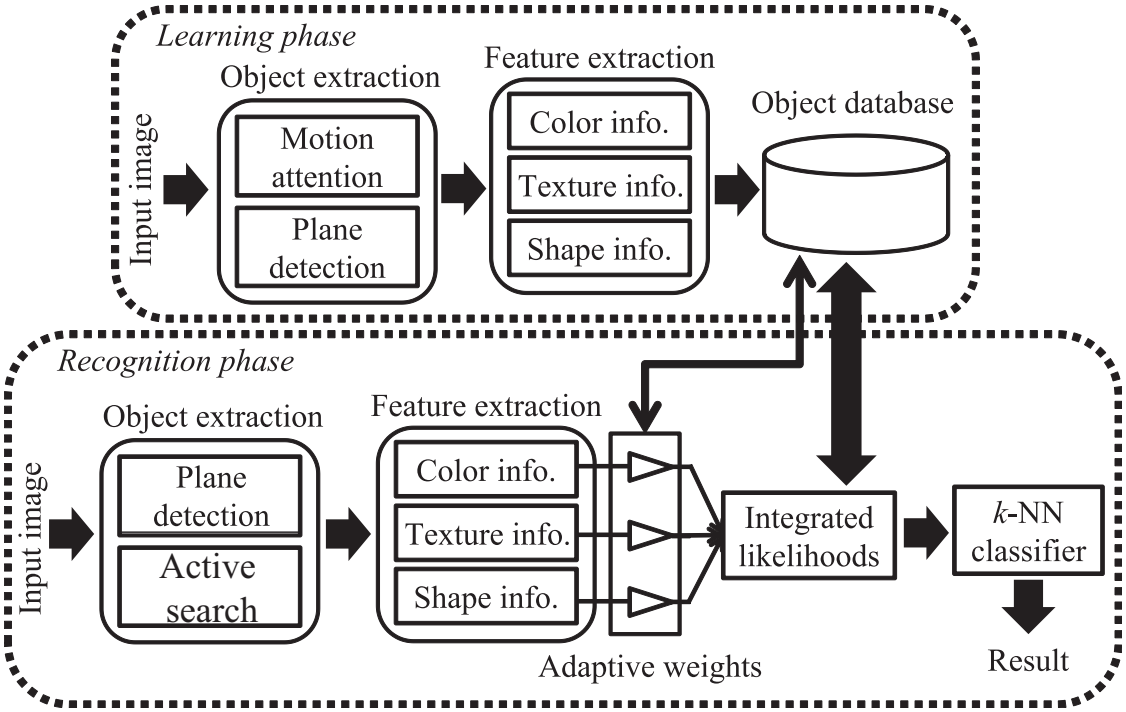

図 2.3: 複数の特徴量を用いた 3 次元物体認識の概要図

重みを調整することで統合する手法を用いる．以下，物体認識手法における各処理を説明する．

まず，動きアテンションによる物体検出手法について述べる．ビジュアルアテンションは人間の選択的注視過程の概念を画像に適用したもので，入力画像中の注目すべき領域を検出するものである．動きアテンションは特に画像中の動きに着目したものであり，画像中の動きに反応して注視点を検出する．本節ではロボットの視野内を動く塊は物体であると仮定することで，シーンからの物体の検出を可能とする．図 2.4 に動きアテンションによる物体検出の概要図を示し，以下にそれぞれの処理について説明する．

処理の第一段階として，キャリブレーションされた RGB 画像を用いて動きの検出を行う．動きの検出手法として計算コストの低いフレーム間差分を用いる．フレーム間差分は前後のフレームにおいて画素値の差分をとることで，第 n フレー

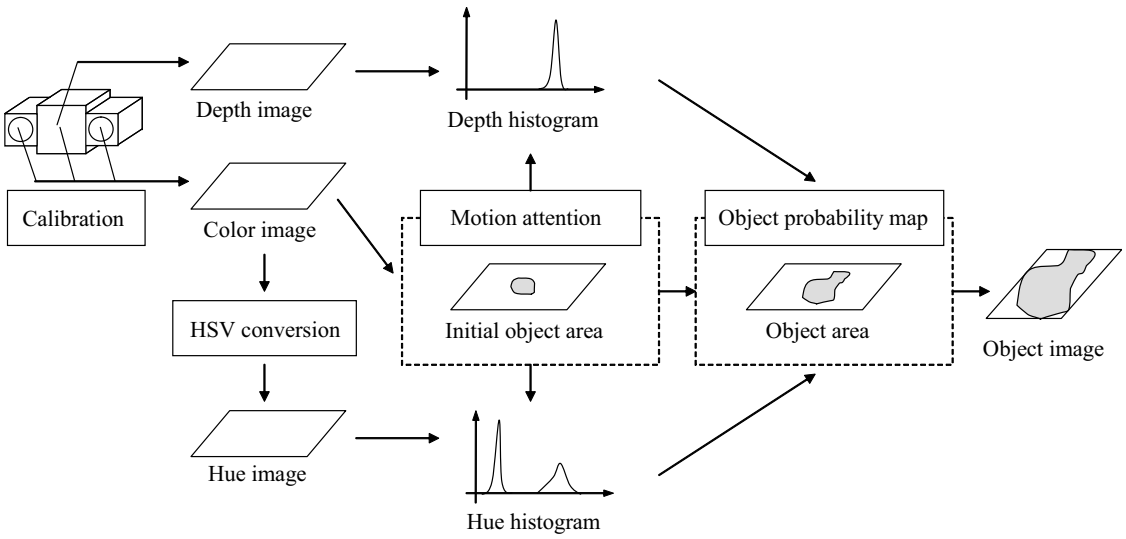

図 2.4: 動きアテンションによる物体検出の概要図

ムでの入力画像を $F_n(u,v)$ とすると，差分画像 $F_n^{\mathrm{diff}}(u,v)$ は

$$F_n^{\mathrm{diff}}(u,v) = |F_n(u,v) - F_{n-1}(u,v)| \tag{2.1}$$

と表わされる．ただし，式（2.1）によって得られる値は入力画像の輝度値による影響が強く，絶対的な数値として評価することはできないため 2 値化処理を施す．ここで，フレーム間差分画像を $\xi \times \xi$ の領域に区切り，各領域における動画素の密度を求める．この密度を表す画像を顕著性マップと呼ぶ．顕著性マップ $P_n^S(u,v)$ は以下の式で計算される．

$$P_n^S(u,v) = \mathrm{LPF}\left[\sum_{i=\xi u}^{\xi(u+1)-1}\sum_{j=\xi v}^{\xi(v+1)-1} F_n^{\mathrm{diff}}(i,j)\right] \tag{2.2}$$

ただし，ξ は整数とし，LPF はローパスフィルタによるフィルタリングを表す．得られた顕著性マップの最大値をとる点が注視点となる．さらに，顕著性マップにクロージング処理（膨張・収縮処理）を行い，注視点からの連結成分を求めることで初期物体領域を抽出する．

この処理によって得られる初期領域は動きのみの領域となっており，CCD カメ

ラより取得した色画像に比べ解像度が低いため，物体の輪郭としては正確さに欠ける．一方，この初期領域には注目する物体が含まれる可能性は高く，領域内の情報を注目物体の情報として利用することは可能であると考えられる．そこで，初期領域中の距離と色相のヒストグラムを利用する．距離のヒストグラムは TOF カメラで取得した距離画像から計算し，色相のヒストグラムは入力画像を HSV 表色系に変換した画像から計算する．これらのヒストグラムを確率密度関数として扱い，距離画像 $G(u,v)$ と色相画像 $H(u,v)$ のそれぞれについて画像中の物体である確率を示す確率マップを作成する．距離の値を d，色相の値を h としたとき，それぞれのヒストグラムが $g(d)$，$h(h)$ と表されるとすると，距離画像に基づく物体確率マップ $P_G(u,v)$，色相画像に基づく物体確率マップ $P_H(u,v)$ は，

$$P_G(u,v) = g(G(u,v)) \tag{2.3}$$

$$P_H(u,v) = h(H(u,v)) \tag{2.4}$$

となる．さらに，この二つの確率マップの重み和をとることにより，最終的な物体確率マップ $P_O(u,v)$ を計算する．

$$P_O(u,v) = \mathrm{LPF}\left[\omega_d \times P_G(u,v) + \omega_h \times P_H(u,v)\right] \tag{2.5}$$

ただし，ω_d，ω_h はそれぞれ距離と色相の重みである．初期物体領域の抽出と同様に，この物体確率マップに対して 2 値化処理，クロージング処理を行い，注視点からの連結成分を求めることで物体を抽出する．

また，入力画像であるキャリブレーション画像は解像度が低いため，キャリブレーション時の対応関係を用いて，元の色画像に戻すことで解像度の高い画像を得ることができる．動きアテンションを利用した物体検出システムにより対象物体を抽出した例を，図 2.5 に示す．学習フェーズでは，ユーザが物体をロボットに提示して学習させるシナリオを考えているため，基本的に物体抽出部は動きアテンションを用いることとする．

次に，平面検出による物体検出手法について説明する．ここで物体は，平面に支えられて存在しているという仮定を基に，物体を検出する．視覚センサによっ

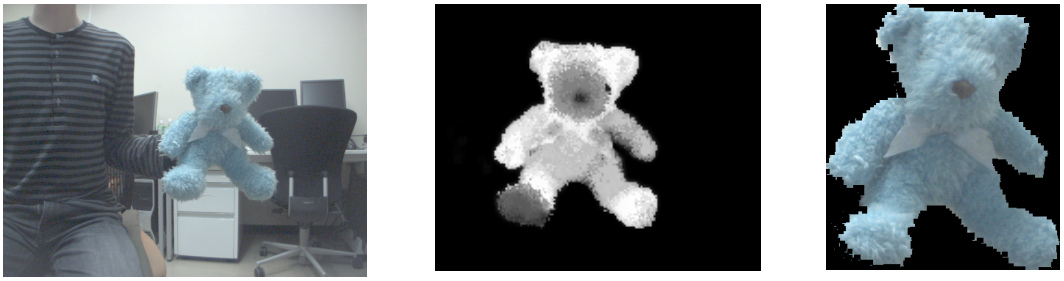

図 2.5: 動きアテンションによる物体検出の例：(a) 入力画像, (b) 物体確率マップ, (c) 抽出された物体

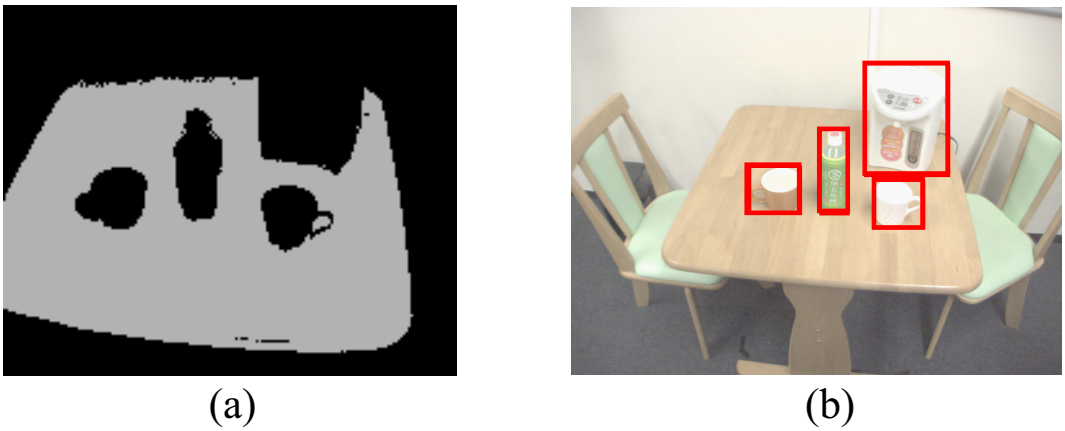

図 2.6: 平面検出による物体検出の例：(a) 検出された平面, (b) 検出された物体

て得られる 3 次元点群にランダマイズドハフ変換 [40] を適用することで，高速に平面を検出することができる．平面抽出後に，全ての距離情報をその平面上に射影し，ラベリングを行うことで物体を検出する．平面検出を用いて物体の検出を行った例を，図 2.6 に示す．机や床といった平面はテクスチャが少ないことが多いため，ステレオカメラを用いた場合，視差が計算できず平面検出に失敗するケースが少なくない．一方，視覚センサを用いることで平面検出の精度が向上し，結果としてロバストな物体の検出が可能となる．平面に基づく物体検出は，基本的には物体を認識する際に用いる．

続いて，3 次元物体認識における学習について述べる．物体を認識するために，

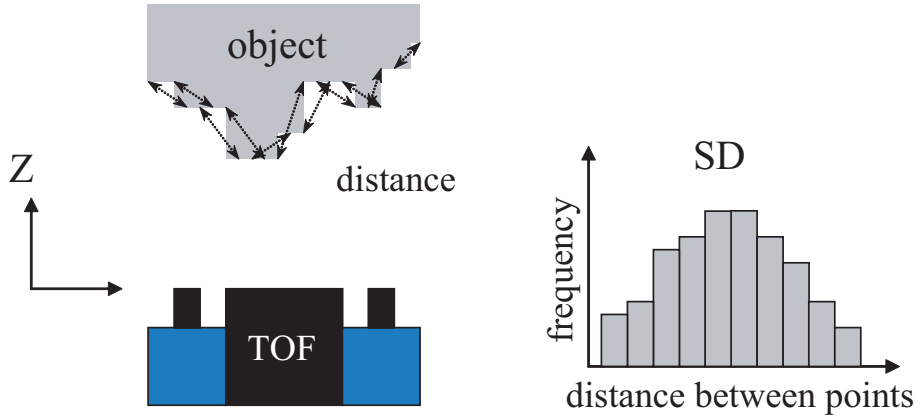

図 2.7: SD の概要図

ロボットは事前に物体を学習する必要がある．この際，人がロボットに物体を様々な方向から見せることで物体学習を行う．また，ロボットが机の上に置いてある物体を把持して様々な方向から観測することも可能である．各物体において，様々な方向からの複数フレーム分に対して検出を行い，物体領域中の特徴量をデータベースに登録する．

ここで，特徴量に対する要件は，スケールと回転，シフト，視点変化に対する不変性である．回転・シフト不変性に関しては，ヒストグラムをベースとした特徴量を用いることで実現可能である．また，スケール不変性に関して，3 次元情報による正規化を行うことで実現できる．視点変化に対する不変性は，特徴量のレベルで実現することが困難であるため，学習時に複数の視点から物体を観測し，それら全ての情報との照合によって解決する．

次に，本節で用いる特徴量 $\tau \in \{color, texture, depth\}$ を説明する．距離情報を用いた特徴量（$depth$）として，Shape Distribution（SD）[41] を用いる．SD は頂点の組み合わせによって様々な特徴を記述できるが，本節では，図 2.7 に示す物体領域中の二つの 3 次元点間の距離に関するヒストグラムを特徴量とする．従来の SD は，3 次元モデル同士の類似度を計るために提案されたものであり，視点に依存しない特徴量であるが，ここでの情報は 2.5 次元であるため視点依存となる．また，従来の SD では 3 次元メッシュからランダムに点群を生成する必要があったが，本節に用いるシステムにおいて 3 次元情報は点群として取得されるため，その必

要はない．SD はスケール，回転，シフトに関して不変性を有するが，全ての点の組み合わせで距離を計算するために計算量が多いという問題がある．この問題は，3 次元点群を間引くことで解決することができる．また，SD では頂点間の距離を計算するため，物体の大きさに関する情報を保持している．従って，形状が同じでも大きさが異なる物体を区別することが可能である．

距離情報を用いた特徴量は，照明条件にロバストであるが，形状が全く同じ物体を識別することができない．そこで，物体表面の色やテクスチャの情報を利用する．照明条件が良い場合には，これらの情報を用いることで認識精度が大幅に向上する．ここでは色情報として，HSV 表色系の H（色相）と S（彩度）のヒストグラムを用いる．具体的には，対象物内の各画素における H と S の値をそれぞれビン数 32（H），10（S）で量子化し，ヒストグラムを計算する．このヒストグラムが色の特徴量（*color*）となる．

本節では，Bag of Keypoints（BoK）[42] をテクスチャ情報として用いる．一般に BoK では，キーポイント数が多い方が認識性能が高くなることが知られている [43]．そこで本節では，キーポイントを密にサンプリングする DSIFT（Dense Scale Invariant Feature Transform）[44] を用いる．これによって得られる特徴量は，回転やスケール変化，照明変化に対してロバストであり，物体を様々な視点から観測する際の特徴量として優れている．ここでは最終的に，DSIFT 記述子をベクトル量子化し，ヒストグラムとして扱う．このヒストグラムはテクスチャの特徴量（*texture*）となり，次のように計算する．学習画像とは異なる画像（室内のランダムなシーン）を複数撮影し，全ての画像から DSIFT 記述子を取得する．これらの特徴量を K 平均法でクラスタリングすることで，500 の代表ベクトル（コードブック）を得る．学習時には，物体領域から計算された DSIFT 記述子を，このコードブックに従ってベクトル量子化する．特徴量は各代表ベクトルの発生回数であり，500 次元のベクトルとなる．

次に，複数の特徴量を用いた物体認識手法について説明する．学習フェーズで作成されたデータベースと，新たに入力された物体の特徴量を比較することで認識を行う．まず最初に，各特徴量 $\tau \in \{color, texture, depth\}$ におけるバタチャリヤ距離を計算する．データベース中の参照物体 o における特徴量 τ のヒストグラ

ムを $\boldsymbol{h}_o^\tau$，認識対象のヒストグラムを $\boldsymbol{h}_{\rm in}^\tau$，次元数を N_τ とすると，バタチャリヤ距離 $D_\tau(\boldsymbol{h}_o^\tau, \boldsymbol{h}_{\rm in}^\tau)$ は以下の式によって求めることができる．

$$D_\tau(\boldsymbol{h}_o^\tau, \boldsymbol{h}_{\rm in}^\tau) = \sqrt{1 - \sum_{i=1}^{N_\tau} \sqrt{\boldsymbol{h}_o^\tau(i) \times \boldsymbol{h}_{\rm in}^\tau(i)}} \tag{2.6}$$

ここで，特徴量 τ における参照物体 o に対する尤度を以下に定義する．

$$P(\boldsymbol{h}_{\rm in}^\tau | o) \propto \exp\left\{ \frac{-D_\tau(\boldsymbol{h}_o^\tau, \boldsymbol{h}_{\rm in}^\tau)^2}{\sigma_\tau^2} \right\} \tag{2.7}$$

ただし，σ_τ^2 は特徴量間の距離のばらつきを調整する係数であり，後で述べるデータベース内の交差検定により事前に求めることとする．全ての特徴量を統合する統合尤度 $P(\boldsymbol{h}_{\rm in}|o)$ は，各特徴量における尤度がそれぞれ独立であると仮定し，以下のように計算する．

$$P(\boldsymbol{h}_{\rm in}|o) = \prod_\tau \{P(\boldsymbol{h}_{\rm in}^\tau|o)\}^{\omega_{\rm in}^\tau} \propto \exp\left\{ -\sum_\tau \frac{\omega_{\rm in}^\tau D_\tau(\boldsymbol{h}_o^\tau, \boldsymbol{h}_{\rm in}^\tau)^2}{\sigma_\tau^2} \right\} \tag{2.8}$$

ただし，$\omega_{\rm in}^\tau$ は各特徴量における重みであり，認識する環境によって適応的に決定する．また，$\sum_\tau \omega_{\rm in}^\tau = 1$ である．従って最終的な認識は，統合距離

$$d(o) = \sum_\tau \frac{\omega_{\rm in}^\tau D_\tau(\boldsymbol{h}_o^\tau, \boldsymbol{h}_{\rm in}^\tau)^2}{\sigma_\tau^2} \tag{2.9}$$

に基づく k 最近傍法によって行う．

本節において，式（2.7）におけるパラメータである σ_τ^2 は，各特徴量における距離のばらつきを表している．このパラメータは，学習フェーズで構築された物体データベース内の交差検定によって決定される．データベースには，物体毎に F フレーム分の色・テクスチャ・距離情報が含まれている．ここで，物体 o のフレーム f の特徴量 τ のヒストグラムを $\boldsymbol{h}_{o,f}^\tau$ とし，集合 $\mathbf{D}_o^\tau$ 内での最小距離を $\min(\mathbf{D}_o^\tau)$

と書くと，分散であるパラメータ σ_τ^2 は，

$$\sigma_\tau^2 = \frac{1}{N_o - 1} \sum_{o=1}^{N_o} (\min(\mathbf{D}_o^\tau) - \mu_\tau)^2 \tag{2.10}$$

によって推定される．ただし，

$$\mathbf{D}_o^\tau = \{D_\tau(\boldsymbol{h}_{o,f}^\tau, \boldsymbol{h}_{o,f'}^\tau) | 0 \leq f < F, 0 \leq f' < F, f' \neq f\} \tag{2.11}$$

である．また μ_τ は，総物体数における $\min(\mathbf{D}_o^\tau)$ の平均であり，N_o はデータベース内の総物体数を表している．

次に物体認識に用いた，特徴量統合における重みの自動決定について述べる．物体認識を行う際にどのような特徴量が有効であるかは，物体の性質（色，テクスチャ，形状など）や認識する環境に依存する．例えば，形状に特徴のあるカラフルで模様の豊富な物体は，色やテクスチャ，形状など，どのような特徴量を用いても比較的認識が容易である．一方，白い食器のような色味のない単色の物体は，色やテクスチャで認識することが困難である．さらに，認識する環境によっても影響を受け，学習時と認識時の環境が大きく異なる場合は，色やテクスチャによる物体認識が難しくなる．特に，白い食器のような物体を暗い中で認識するためには，形状の情報が重要な役割を果たす．本節ではこうした問題に対処するために，式（2.9）の統合重み ω_{in}^τ を適応的に決定する．統合重みを調整するための指針は，次の 2 点である．

1. ある特徴量においてデータベースとの最小距離が大きい場合，その特徴量があまり有効に機能せず，結果として統合尤度を下げてしまい，誤認識につながる可能性がある．つまり入力した特徴量に対する重みは，データベースとの最小距離に反比例させることで，学習時の環境と大きく異なり有効に機能していない可能性のある特徴量を無効化する．

2. データベースに含まれる複数の物体が非常に近い特徴量を持つ場合，その特徴量はそれらの物体を判別する際に有効に機能しないことになる．従って，入力した特徴量に対する重みは，最も距離の近い物体とは異なる物体の中で

最も近い距離に比例させることで，データベース内で類似の特徴を持つ物体の影響を考慮する．

以上の点を考慮し，最小距離 $\min(\mathbf{D}^{\tau}_{\mathrm{in}})$ と，それとは異なる物体との最小距離 $\min(\hat{\mathbf{D}}^{\tau}_{\mathrm{in}})$ を用いて，重み $\omega^{\tau}_{\mathrm{in}}$ を次式の様に定義する．

$$\omega^{\tau}_{\mathrm{in}} = \frac{\gamma_{\tau}}{\sum_{\tau}\gamma_{\tau}}, \quad \gamma_{\tau} = \frac{\min(\hat{\mathbf{D}}^{\tau}_{\mathrm{in}})}{\min(\mathbf{D}^{\tau}_{\mathrm{in}})} \tag{2.12}$$

ただし，

$$\mathbf{D}^{\tau}_{\mathrm{in}} = \{D(\boldsymbol{h}^{\tau}_{\mathrm{in}}, \boldsymbol{h}^{\tau}_{o,f}) | 0 \leq o < N_o, 0 \leq f < F\} \tag{2.13}$$

$$\hat{\mathbf{D}}^{\tau}_{\mathrm{in}} = \{D(\boldsymbol{h}^{\tau}_{\mathrm{in}}, \boldsymbol{h}^{\tau}_{o,f}) | 0 \leq o < N_o, o \neq o^{\tau}_{\mathrm{min}}, 0 \leq f < F\} \tag{2.14}$$

である．また，o^{τ}_{min} は距離が最小となる物体のインデックスであり，$o^{\tau}_{\mathrm{min}} = \underset{o}{\mathrm{argmin}}\, \mathbf{D}^{\tau}_{\mathrm{in}}$ である．

2.3.5 近赤外線反射強度を用いた材質認識

ここではまず，材質認識に用いる特徴量について議論する．材質情報は各材質によって異なるため特徴量として重要である．本節では，材質情報は反射係数より抽出する．また，よりロバストな認識を行うために距離や入射角を考慮する．

本節では，TOF カメラから取得できる近赤外線（NIR : Near Infrared）反射強度を利用した材質認識手法を提案する．文献 [45–47] において，TOF カメラは照明条件に影響されないと報告されている．しかし，近赤外線反射強度の輝度はインテグレーションタイム δt，物体までの距離 d と入射角 φ によって大きく異なる．本節では近赤外線反射強度の性質を理解するために，放射輝度モデルを考慮する．ここでは物体の表面における反射を拡散反射として考え，近赤外線反射強度に対して図 2.8 に示す放射輝度モデルを適用する．ある一定の波長 λ（TOF では $\lambda = 850nm$）において放射輝度モデルは以下のように表すことができる．

$$\text{Ł}_{Q}(\lambda) = \int I_e(\lambda, d)\rho_d(\lambda)cos\varphi dt = I_e(\lambda, d, \delta t)\rho_d(\lambda)cos\varphi \tag{2.15}$$

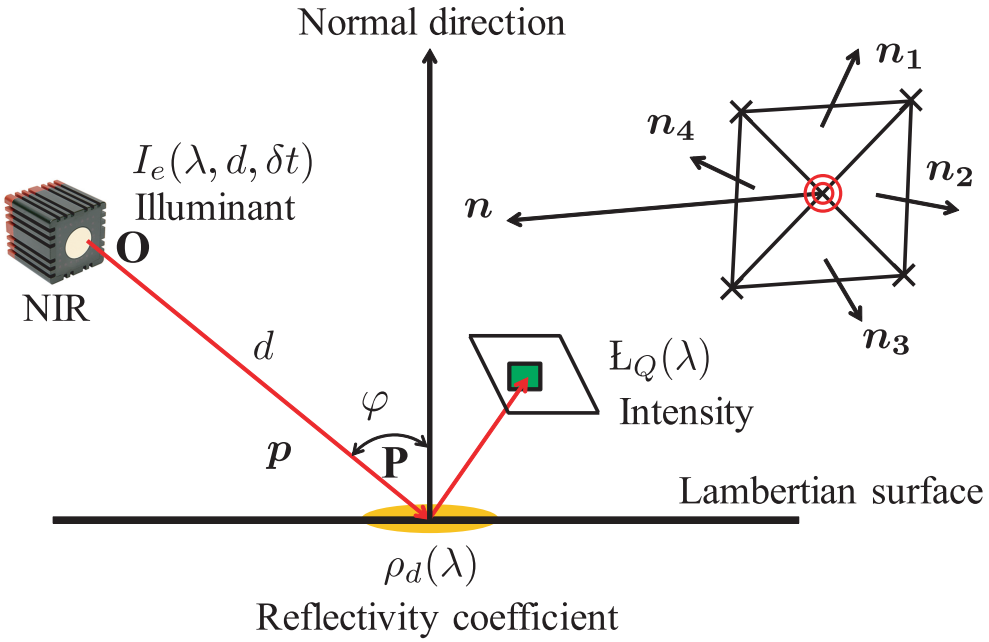

図 2.8: 放射輝度モデルを用いた反射係数

ここで，$Ł_Q$，I_e，ρ_d，φ，d，δt はそれぞれ放射輝度，光源（TOF カメラからの近赤外線），反射係数，光源（カメラの中心）に対する入射角，光源から物体までの距離，インテグレーションタイムを表している．反射係数は材質によって異なるため，材質情報に相当するものと考えられる．法線ベクトル $\boldsymbol{n}$ と入射角 φ は以下の式より求められる．

$$\boldsymbol{n} = \frac{\sum_{k=1}^{N} \omega_k \boldsymbol{n}_k}{\sum_{k=1}^{N} \omega_k}, \quad \cos\varphi = \frac{\langle \boldsymbol{p}, \boldsymbol{n} \rangle}{\|\boldsymbol{p}\| \|\boldsymbol{n}\|} \tag{2.16}$$

ただし，$\boldsymbol{p}$ は TOF カメラの中心点 O と反射面上の点 P における位置ベクトルであり，ω_k は法線ベクトルの近傍法線における重みである．重み ω_k は近傍法線ベクトルの全ての 3 次元点が存在する場合 1 とし，そうでない場合 0 とする．従って，近赤外線反射強度は材質，距離，入射角によって異なり，材質情報を $\boldsymbol{m}(Ł_Q, \delta t, d, \varphi)$ とする．これらのパラメータは TOF カメラより取得可能である．これらの情報を利用し，図 2.9 に示す材質認識手法を用いる．

まず，SVM（Support Vector Machine）を用いた材質認識について説明する．既

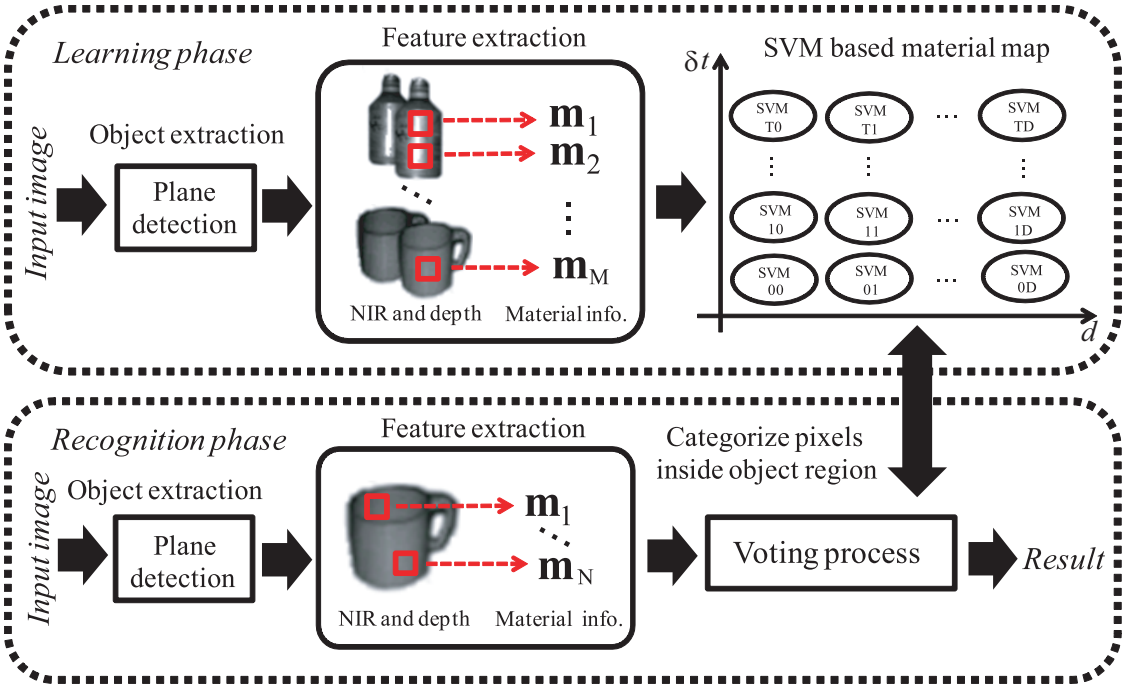

図 2.9: 材質認識の概要図

に述べたように，特徴ベクトル $\boldsymbol{m}(Ł_Q, \delta t, d, \varphi)$ は物体の材質の反射係数に依存する．よりロバストな材質認識を行うために，これらの情報を様々な組合せで用いる．しかし，全ての情報を一つの識別器に組み込むと，学習時間や使用メモリの増大といった問題が生じる可能性がある．そこで，図 2.9 のように複数の SVM を用いて，材質マップを生成する．一般に，SVM は 2 クラスの分類に用いられることが多く，識別器として高性能である．また，多クラスに対して 2 クラスの組み合せを全て考慮することで，多クラスの分類問題にも応用できる．本節では，SVM の実装として LIBSVM [48] を用いる．ここでは，物体領域内の各画素の入射角 φ と近赤外線反射強度 $Ł_Q$ の 2 次元のデータを使用し SVM の学習を行う．

材質マップを作成するために，近赤外線反射強度画像及び距離画像を学習フェーズにおいて取得しデータベースを構築する．ここで，対象となる画素 $\boldsymbol{x}^m = \{x_1^m, x_2^m, \ldots, x_{N_m}^m\}(d^{\min} \leq d(\boldsymbol{x}^m) \leq d^{\max}, \delta t^{\min} \leq \delta t(\boldsymbol{x}^m) \leq \delta t^{\max})$ を各材質 $m \in \{1, 2, \ldots, M\}$ に対して，データベースの中から選択する．ただし，$d^{\min}$, $d^{\max}$, $\delta t^{\min}$,

$\delta t^{\max}$ は，データベース内のそれぞれ最小と最大となる距離 d 及びインテグレーションタイム δt であり，T と D はそれぞれ，δt 及び d のグリッド数を表している．また，$d(*)$ と $\delta t(*)$ の表記はそれぞれ，画素 $*$ の距離とインテグレーションタイムであり，N_m は材質 m となる画素数を表している．これにより，$\boldsymbol{x}^m$ の材質情報 $\boldsymbol{m}(\text{Ł}_Q, \delta t, d, \varphi)$ が決定される．図 2.9 に示す材質マップ $(d, \delta t)$ を構築するために，$\boldsymbol{x}^m$ 内の画素がグリッド化された距離とインテグレーションタイム (n_d^m, n_t^m) によって分類される．i 番目の画素 $x_i^m \in \boldsymbol{x}^m$ に対して, n_d^m と n_t^m はそれぞれ以下のように計算する．

$$n_d^m = \left\lfloor \frac{d(x_i^m) - d^{\min}}{(d^{\max} - d^{\min})D} \right\rfloor, \tag{2.17}$$

$$n_t^m = \left\lfloor \frac{\delta t(x_i^m) - \delta t^{\min}}{(\delta t^{\max} - \delta t^{\min})T} \right\rfloor, \tag{2.18}$$

これにより，各グリッド (n_d^m, n_t^m) は画素 $\boldsymbol{x}^{mdt} = \{x_1^m, x_2^m, \ldots, x_{N_{mdt}}^m\}$ を持つようになる．ここで，mdt は距離 d，インテグレーションタイム t となる画素であり，材質 m に属する．ただし，N_{mdt} は画素 mdt の数を表している．学習フェーズの最終段階として，$\boldsymbol{x}^{mdt}(m \in \{1, 2, \ldots M\})$ を多クラス SVM を用いて学習することでグリッド (n_d^m, n_t^m) のモデルを作成する．

認識フェーズにおいて，平面検出を用いて対象物を抽出し，材質情報を計算する．δt と d が既知であれば，材質マップを利用することができる．ここで，δt は TOF カメラのパラメータを調整することで設定できるが，d は TOF カメラより取得可能である．画素の d と δt が分かれば，グリッド (n_d^m, n_t^m) は式（2.17）及び（2.18）が決定され，そのグリッドに対する SVM モデルより画素の材質を推定することが可能である．

上述の手法を用いることで，画素レベルの結果が得られるが，信頼性は低い．そこで，物体領域内の全ての画素に対して結果を考慮する物体ベースのスムージングを導入する．ここで M を，材質のカテゴリ数とする．V_i^m は i 番目の物体領域内 D_i における材質 m に対する投票数を表している．各投票 $v_i^m \in \{1, 2, \ldots, M\}$ は前節で述べたように選択された SVM モデルより予測されたラベルとなる．物体領

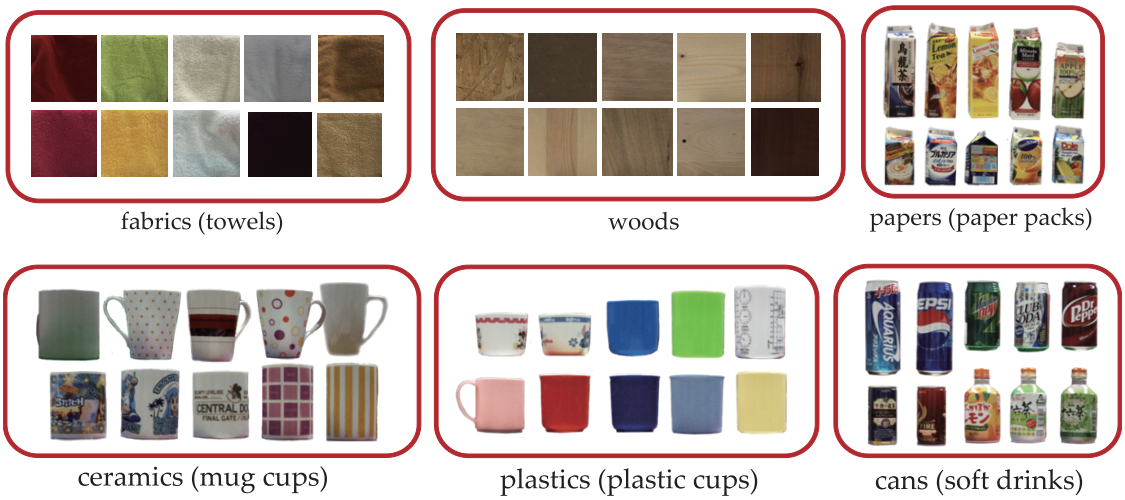

図 2.10: 材質認識実験に用いる物体

域内の画素 (u, v) において，材質 m の確率 $P(u, v|m, i)$ は次のように計算することができる．

$$P(u, v|m, i) = \frac{V_i^m}{\sum_{k=1}^{M} V_i^m} \tag{2.19}$$

この確率は画素 (u, v) における材質認識の信頼度を表しており，最終結果は投票の最多として求める．物体ベースのスムージングでは，対象物の材質を決定するために物体領域内の画素に対する最終の投票処理が行われる．最終的に，物体の材質 $\hat{m}_i$ は次の式より算出する．

$$\hat{m}_i = \underset{m}{\text{argmax}} \sum_{(u,v)\in D_i} P(u, v|m, i) \tag{2.20}$$

本節に用いる材質認識はロボットに実装されるため，認識結果の信頼度が低い場合は，ロボットが自身の身体性を利用した再認識を行うことが可能であると考える．つまり，物体を把持して学習のときと同じ角度と距離で認識を行うことで認識結果の信頼性を高めることができる．

次に，材質認識の評価実験について述べる．本節に用いる材質認識の精度を測

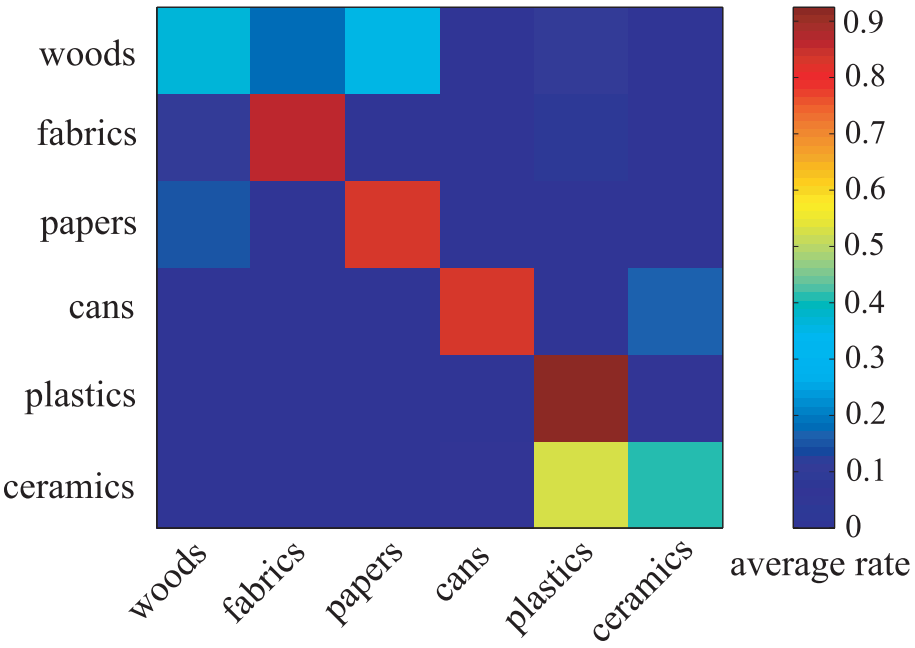

図 2.11: 材質認識の混同行列

るために，図 2.10 に示す 6 クラス（木，布，紙，プラスチック，缶，セラミック）に分けられる 60 個の物体を用いて実験を行った．60–80cm の距離で Leave-one-out 法による認識実験を行った．平均認識率の混同行列は図 2.11 に示した．全ての距離における平均認識率は 70.2%であった．誤認識として多かったのは，木材とセラミックのクラスであった．これらのクラスに対する近赤外反射強度のばらつきが大きいことが原因として考えられる．

一方で，これらの材質を可燃ごみ（木，布，紙）及び不燃ごみ（プラスチック，缶，セラミック）として分けた場合，97.7%の認識結果が得られた．掃除タスクでは，これらの結果を用いることで，ごみを燃・不燃で分別することが可能である．また，実際のシーンでの材質認識結果の例を図 2.12 に示す．布と缶，およびプラスチックカップをシーンの中から抽出しその材質が正しく認識されていることが分かる．各クラスの認識信頼度も確率マップとしてプロットした．

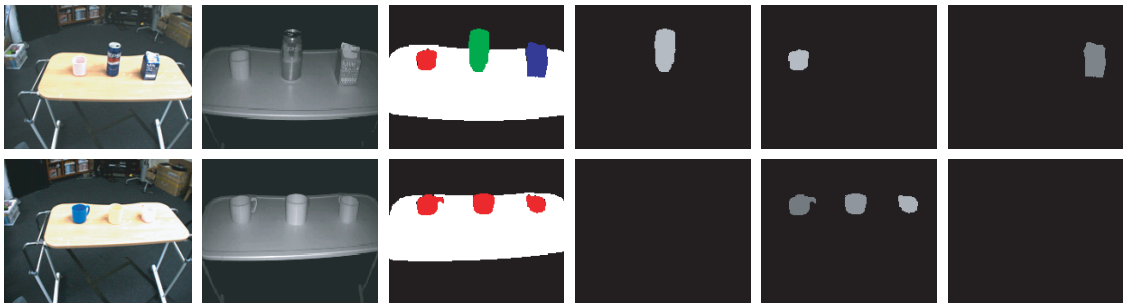

図 2.12: 実際のシーンでの材質認識の例：(a) 色画像（1024 × 768），(b) 近赤外線反射強度画像（176 × 144），(c) 分割画像（176 × 144），(d) 缶の確率マップ（176 × 144），(e) プラスチックの確率マップ（176 × 144），(f) 紙パックの確率マップ（176 × 144）．分割画像において，白い画素は机を表しており，黒い画素は精度の低い距離画素を表現する

2.3.6 GMM を用いた細かい物体の検出

机の上のペットボトルや缶など，高さのある物体をロボットに片付けさせたい場合，平面検出を利用した物体検出と前述の認識システムによって認識し把持することができる．一方，紙くずなど小さく高さのないごみは，平面検出やアクティブ探索による検出が困難である．ここでは，机の表面の色を事前に学習しその色との差異によって検出する．具体的には，机の色（HSV 色空間の H と S）をガウス混合モデル（GMM : Gaussian Mixture Model）によってモデル化し，この GMM を用いて机とごみを区別する．しかし色だけでは，机とごみの色が似ている場合，区別することが困難である．そこで，材質情報を含む近赤外線反射強度を利用する．机の色の学習時に近赤外線の反射強度も同様に GMM を用いてモデル化し，これにより表面のごみを検出する．ここでは，材質情報の入射角 φ は学習フェーズと同じになるように調整可能なため，無視することができる．従って，卓上の色（H と S）と近赤外線反射強度 L_Q からなる特徴ベクトル $\boldsymbol{I}(H, S, \mathrm{L}_Q)$ を GMM でモデル化することで，よりロバストな細かい物体検出を行える．細かい物体検出の概要図を図 2.13 に示す．

学習フェーズにおいて，机に所属する全画素の尤度の最大化を行う．卓上の画

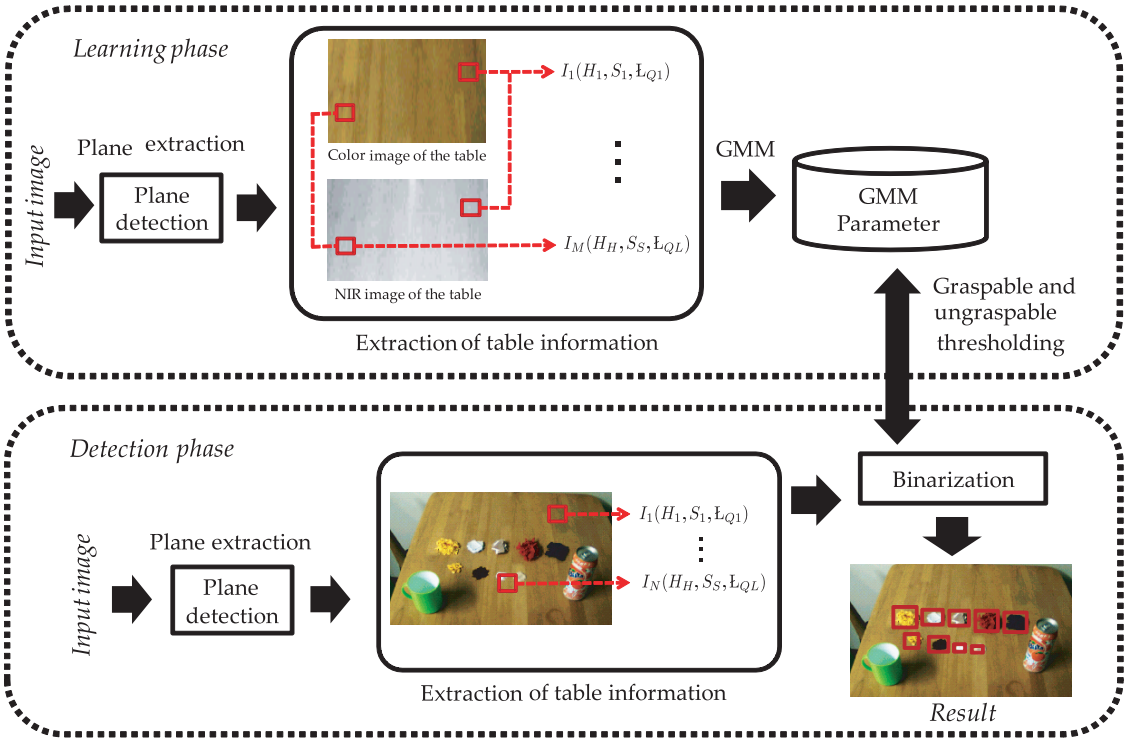

図 2.13: GMM を用いた細かい物体検出の概要図

素 i の特徴ベクトルを $\boldsymbol{I}_i$ とし，K を混合モデル数からなる GMM のパラメータ $\Theta = (\pi_1, \mu_1, \Sigma_1, \ldots, \pi_K, \mu_K, \Sigma_K)$ とする．特徴ベクトル $\boldsymbol{I}_i$ の尤度は次のように求めることができる．

$$P(\boldsymbol{I}_i|\Theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{I}_i|\mu_k, \Sigma_k), \tag{2.21}$$

ただし，$\mathcal{N}(*|*)$ はガウス分布を表す．パラメータ Θ は EM アルゴリズムを用いて推定することができる．本節では，GMM における混合モデル数は経験的に決め，$K = 3$ とする．今後は，Dirichlet Process Mixture（DPM）[49, 50] を用いることで混合モデル数を自動的に推定することができる．

認識フェースでは，卓上における各画素 $\boldsymbol{I}_i$ の尤度を式（2.21）より算出し，次

式を用いることで 2 値化を行い，2 値化マップを作成する．

$$b(\boldsymbol{I}_i) = \begin{cases} 0 & (P(\boldsymbol{I}_i|\Theta) \geq P^t) \\ 1 & (P(\boldsymbol{I}_i|\Theta) < P^t) \end{cases} \tag{2.22}$$

ただし，P^t は閾値であり，経験的に決定する．次に，2 値化マップに対して膨張･収縮を行う．最終的に，ラベリング処理を行い，候補領域 $S(R^\ell)$ $(\ell \in \{1, 2, \ldots, N_\ell\})$ の面積 R^ℓ を次式より計算する．

$$S(R^\ell) = \sum_{i \in R^\ell} b(\boldsymbol{I}_i), \tag{2.23}$$

ただし，N_ℓ はラベル数を表している．候補領域をフィルタリングし，条件 $R^\ell \in \{S^{\min} \leq S(R^\ell) \leq S^{\max}\}$ を満たす領域のみ検出される物体として選択する．ここで，$S^{\min}$ 及び $S^{\max}$ はそれぞれ，許可される最小面積と最大面積を表しており，それらの値は経験的に決定される．こうして検出されたごみ（選択された各領域 R^*）は把持不可能なため，ロボットはハンディ掃除機を利用してごみを掃除する．

次に，細かい物体検出の評価実験について説明する．本節で用いる細かい物体検出の精度を測るために，図 2.14 に示すように紙くずや粉（砂糖，塩，コーヒー）など 15 個の物体を用いて 4 種類の机で行った．ここで，各机において 2 種類のシーン（正常のシーンと混雑しているシーン）を取得し，合計 116 個の細かい物体を含んだ 12 シーンで実験を行った．本節では，PASCAL Visual Object Challenge（VOC）評価法 [51] を用いて検出率を求めた．この評価法では，検出候補の領域が正解領域の半分以上に含まれれば正解とする．また，複数の検出結果が同じ一つの正解を指すのであれば，一つが正解と認められ，それ以外を誤検出として扱う．結果として，79.3%の細かい物体検出率が得られた．近赤外線反射強度が導入されたため，照明変化に対する悪影響を防ぐことができた．実際，コーヒーなどのような細かい物体の近赤外線反射強度の値は机のものに近かった．このような細かい物体において，白い机のように机の色と異なる場合，色差があるため検出することができた．本節における把持不可能な物体検出の主な目的が，掃除機のホースを動かすための大雑把な位置の特定であることを考慮すれば，これは十分な精度であると考えられる．

図 2.14: 細かい物体検出実験に用いる物体

2.4 掃除タスクの実現

前述の視覚認識システムを用いて，2.2 節で述べた掃除タスクを実現する．ただし，移動や物体の把持は LRF やカメラの 3 次元情報を用いて実現できているものとする．実装する掃除タスクの流れを，図 2.15 に示し，タスクの詳細は次のようになる．

このタスクにおいてロボットはまず，対象となる机の上のきれいな状態を学習しておく必要がある．ロボットは机の上にあるべき物体を認識・記憶する．さらに，2.3.6 で述べた GMM のパラメータを推定することで，机の上の色と材質を学習する．このパラメータが学習されると，ロボットはきれいな状態に関する情報を学習する．ここで，把持可能な商品は平面検出より検出され，3 次元物体認識を行うことで商品を認識しそれらの位置を記憶する．卓上と商品を記憶した後は，掃除タスクをいつでも行うことができる．タスクの実行は，視覚認識システム（2.3.2 節を参照されたい）による卓上の認識と，プランニングと実行に分けられる．プランニングと実行にはロボットの移動と把持可能・不可能な物体の操作が含まれる．

タスク開始時に，ロボットは対象の机に移動する．LRF を用いているため，iterative-closest-point（ICP）[52] に基づく自己位置推定と地図作成を行うことができる．自律移動には，Rapidly-exploring random tree（RRT）[38] に基づいたパスプランニ

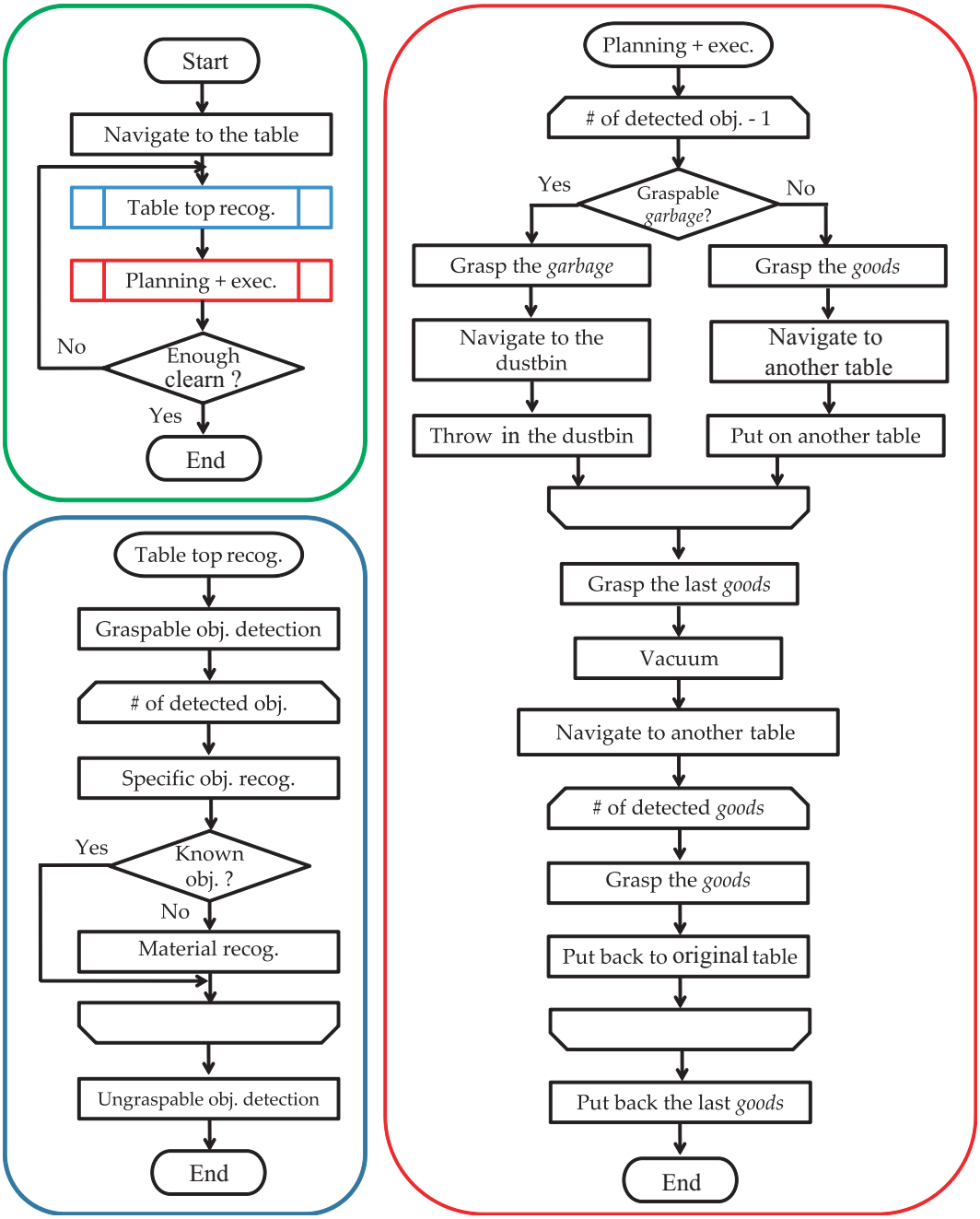

図 2.15: 掃除タスクの流れ：緑色のブロックは全体タスクを表し，青色と赤色のブロックはそれぞれ，卓上の認識，ロボットのプランニングと実行の詳細を示す

ングが用いられる．対象の机に到達した後，把持可能な物体の検出（平面検出を用いるかアクティブ探索を用いる）を行う．検出された物体が既知（商品）であれば，右腕で掃除機を把持するため左腕で掴んで別の机に運ぶ．一方検出された物体が未知（ごみ）であれば，そのごみの材質を認識し不燃か可燃かを判別する．把持できるごみが複数ある場合，その都度ごみ箱に捨てに行く．商品に対しても同じように，その都度別の机に運ぶ．物体操作のパスプランニングには，RRT アルゴリズムを用いる．

把持可能なごみを掃除した後，把持できない細かい物体検出を GMM に基づく物体検出によって行う．細かい物体が検出された場合，XBee 無線モジュールを介して自動的に掃除機をオンにする．次に，掃除機のホースを検出された細かいごみの領域（2.3.6 節を参照されたい）に移動させ掃除する．より効率的に掃除を行うために，掃除機を検出された細かいごみの中心辺りに移動させる．細かいごみが検出されなくなった段階で，掃除機を自動的にオフにする．最終的に，把持可能な商品をもとの机に戻せば掃除タスクが終了する．

2.5 タスクの実行結果と議論

ここでは，実際に掃除タスクを実装したロボットによる掃除タスクの実行結果について述べ，その結果を通してタスクと概念・理解の関係性について議論する．

2.5.1 掃除タスクの評価

本章で提案する掃除タスクを，実際のロボットプラットフォームに実装した．図 2.16 には，いくつかの机の上に対して実際に視覚認識を行った結果を示す．この図における認識結果は，捨てる把持可能な物体（ごみ）が緑色の矩形で示され，別の机に運ぶ物体（商品）と掃除機で吸う細かい物体（把持不可能なごみ）がそれぞれ青色と赤色の矩形で描画されている．この結果より，十分な精度で視覚認識システムが機能していることが分かる．

こうした視覚認識の結果に基づき，ロボットは行動を計画し，実際に掃除タスクを実行することになる．図 2.17 に，実際にロボットが図 2.15 の流れ（2.4 節を

図 2.16: 机の上の認識結果．各机に対して，上段が色画像（1024 × 768），中段が距離画像（176 × 144），下段が近赤外線反射強度画像（176 × 144）．検出結果：掃除機で吸うべきごみ（赤色の枠），特定物体認識によって認識された物体（青色の枠），材質認識によって材質が特定された物体（緑色の枠）

参照されたい）に従って掃除タスクを実行している様子を示す．

ロボットは卓上の初期状態を記憶した後，対象となる机に移動し認識する．その後，把持可能なごみを探索しそれをごみ箱に捨てる．全ての把持可能なごみが無くなれば，把持可能な商品を一個になるまで別の机まで運ぶ．残り一個の商品を左手で掴んで，把持不可能なごみを右手の掃除機で吸う．把持不可能なごみが無くなれば，商品が置かれた机に移動しそれらをもとの位置まで運ぶ．

掃除タスクを評価するために，異なる机の状態に対してタスクを 10 回行った．その 10 回の試行の内，7 回成功して 3 回失敗した．失敗した例として，ロボットがごみをごみ箱に捨て損ねたケースや，商品をもとの机に置くときに大きなずれが生じたケースなどが挙げられる．この場合，対象となる机はきれいになったが，掃除全体としては望まれない結果となっている．成功した試行において，主観的

図 2.17: 掃除タスクの実行例．タスクは図 2.15 に示す流れに従って行う．ロボットの主な行動として，移動，認識，把持不可能なごみの掃除

にきれいとなったものは 5 回であり，残りの 2 回は把持不可能なごみを机に残した．掃除タスク終了後のいくつかの状況を図 2.18 に示す．全体的に，商品をもとの位置に置く精度，及び掃除機を把持不可能なごみの位置に置く精度が低い．これらの原因として，ロボットの移動誤差が考えられる．

2.5.2 議論

ここで実装した掃除タスクの問題は，その精度の低さが本質ではない．なぜなら，精度の問題は，例えばごみ箱にマーカーを貼ることなどで大幅に改善することができ，その後のこうした改善で最終的にはほとんど失敗するケースを無くすことができることを確認している．従って，食後のテーブルを片づけるウェイター

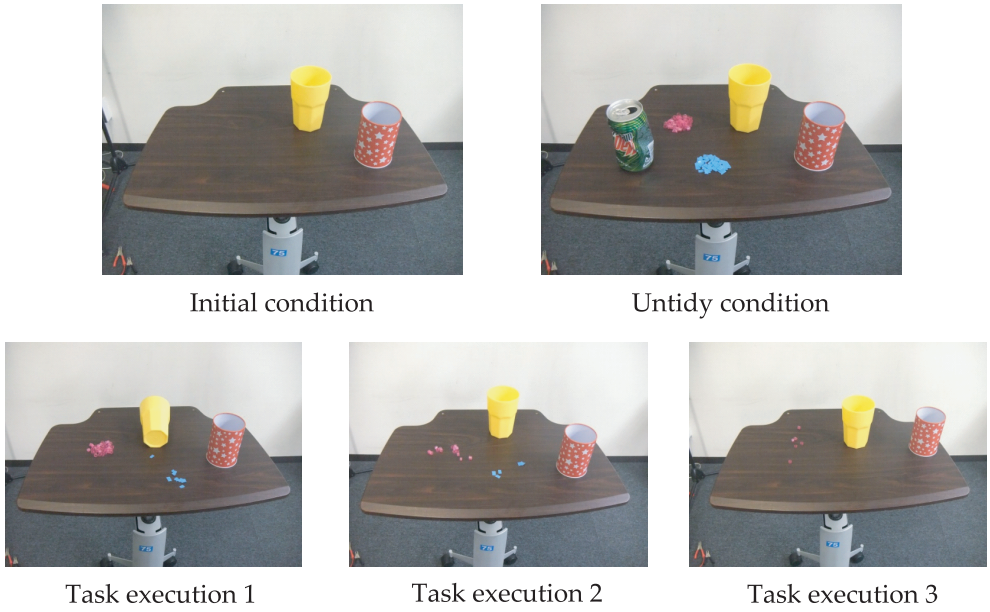

図 2.18: 掃除タスクが終了した状態の例：初期（きれいな）状態（左上），汚い状態（右上），タスクが終了した状態（下）

のようなロボットとしては，非常に優れたものを実現可能である．しかし依然として，掃除のできるロボットとしては重大な問題をはらんでいることは明らかであろう．この問題を議論するためには，そもそも掃除するということがどのようなことなのかを考える必要がある．著者の実装したロボットには何が足りなかったのか？この問いに答えることで，ロボットの知能に対して新たな局面を生むのではないかと期待する．

「掃除とはどいうことか，何のためにするか」という問いかけに対して，一般的には「汚いところをきれいにする」という答えが多いであろう．つまり,「汚い」や「きれい」という状態の判断と「汚い状態」を「きれいな状態」にするための行動決定の問題であると言える.「汚い」や「きれい」という概念については，感性の問題であり本論文のスコープから外れるため後に議論することとし，ここではそうした状態が判断できるとしてどのように行動すればよいかという問題を考えてみる．行動を決定するためには，どのような時にどのような行動をすべきかということを経験的に知っていなければならない．

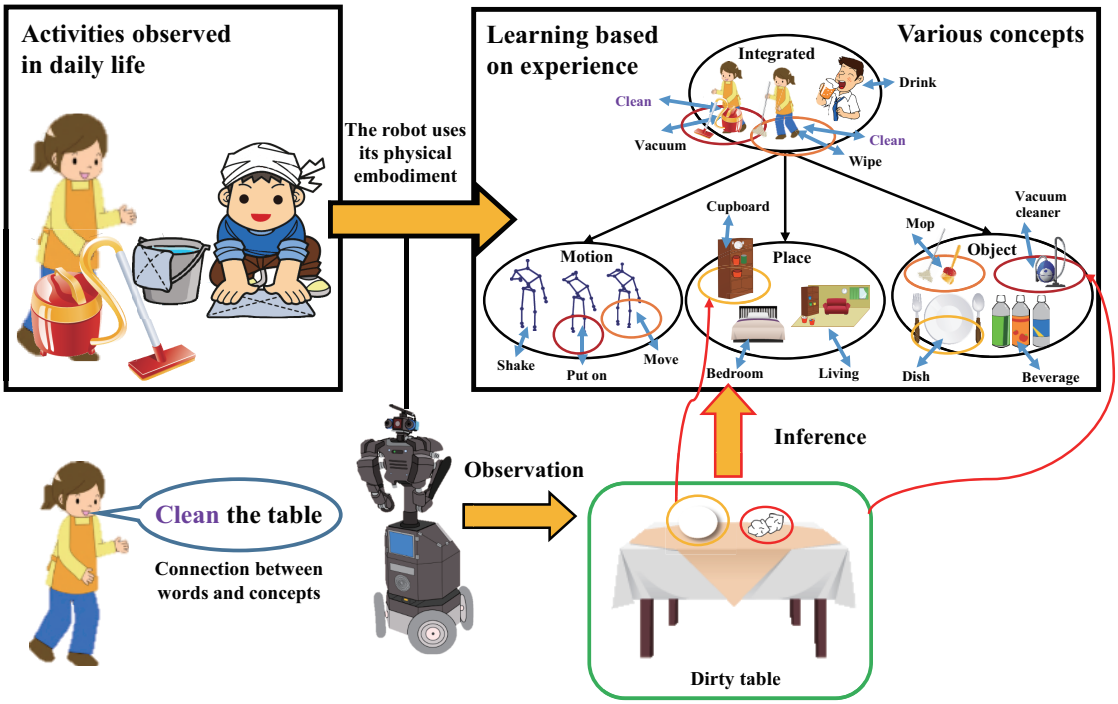

図 2.19: 多様な概念を用いた掃除の概要図

例えば，掃除機を持って細かいごみの上で動かすことでごみを取り除くことを経験すると，掃除機を手で持って動かすという「掃除機をかける」という「行為」が，物体の概念と動きの概念の関係性において表現されることになる．また,「雑巾」を手で左右に動かすという動きの関係性が,「拭く」という行為の概念を表現する．このような行為の概念がさらに組み合わさって，掃除という統合的な概念を形成していると考えることができる．こうした概念は，図 2.19 に示すように現在の状況，つまりは知覚情報によって得られる文脈によって適切なものが駆動される．例えば，机の上にある紙くずなど細かいごみを観測すると，紙くずと関係する多様な概念が発火し，これによって紙くずの上で掃除機をかければそれらをなくすことが可能であると推論することができるようになる． さらには，常識と一般に呼ばれるような背景知識も必要であろう．例えば，皿は食器棚にあるべきであるといった物体概念と場所の概念との関係性である．つまりこうした背景知識は，図 2.19 に示すように概念間の関係を表現する構造の中に埋め込まれている

と考えることができる．こうした概念間の階層的で複雑な関係を構造化することが本質的な問題解決につながるのではないかと考える．

言語についても同様に考えることができる．既に述べたように，語彙は各概念の音韻ラベルとして考えることができる．従って，概念の階層構造においても，各概念において適切な音韻ラベルが結び付くことになり，これらの概念を言語によって駆動することができるであろう．ユーザーの「掃除しろ」という命令の解釈は，まさに上記の掃除という上位概念を駆動し，現在の知覚情報と共に駆動される物体概念，動き概念，場所概念とそれらの関係性に基づく予測によって適切な行動を選択することに他ならない（図 2.19）．これがまさに，著者が目指す理解の仕組みである．ただし，実際に掃除を行うためには，物理的な制約や時間的順序などより複雑なプランニングが必要となることは明らかである．この問題は，概念をどのように利用するかという問題である．一方，本論文で主に扱うのは，概念構造をどのように自律的に獲得できるかという問題であると言える．従って，獲得した概念の活用については後の議論としたい．

「掃除」とは，多様な概念の階層的な相互依存関係から構成される概念である．こうした多様な概念の形成と，それらの階層的な構造の構築こそがロボットの知能として求められるに違いない．これは，掃除に限った話でないことは明らかであり，知能一般に拡張できるものであると考える．本論文では，次章以降でこうした多様で階層的な構造を持つ概念を，ロボットがいかに自律的に獲得するかについて議論する．

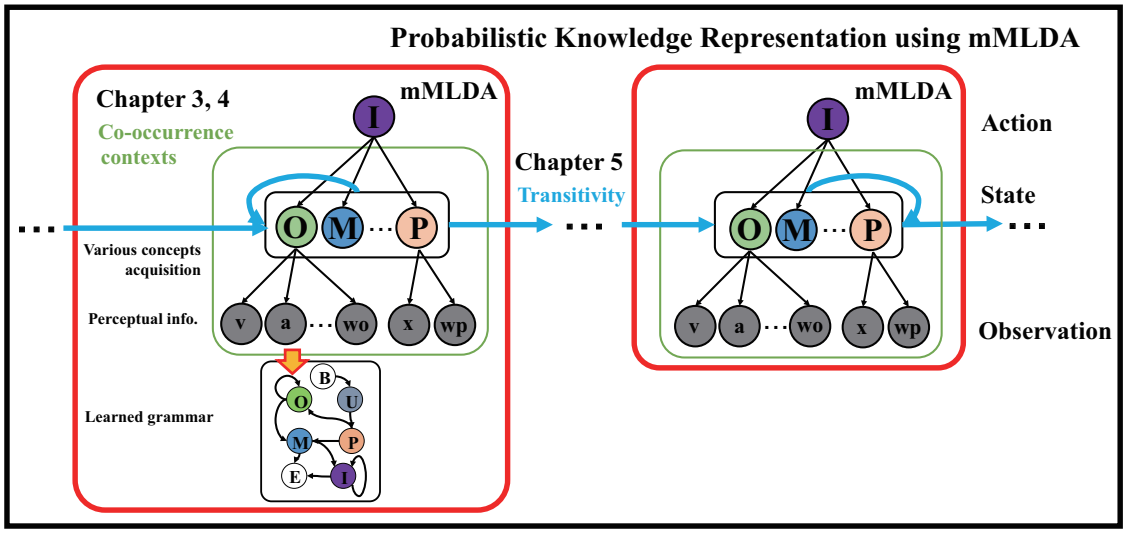

図 2.20: mMLDA を用いた確率的知識表現

2.6 まとめ

本章では，ロボットが家庭環境で活動することを前提として，掃除タスクを取り上げて概念や言語理解の必要性について議論した．まず RoboCup@Home のタスクを実現するために，著者はヒューマノイドロボットによる掃除タスクを実装し，そのための視覚認識システム，プランニング・実行について述べた．掃除タスクをより精度良く実行するための手法として，階層構造の一貫性を利用する視覚認識システム [53] や環境の曖昧性やロボットの身体性を考慮したプランニング [54] が挙げられる．しかし，人のように掃除することのできるロボットとして未だ重大な問題をはらんでいることは明らかであり，ここではその問題や解決方法について述べた．著者は,「掃除」とは多様な概念の階層的な相互依存関係から構成される概念であると捉え，こうした多様な概念の形成と，それらの階層的な構造の構築こそがロボットの知能として重要であると考えている．

ここで，掃除タスクのみならず一般のタスクも考慮した上で，ロボットがタスクを遂行するために必要な知識をどのように表現し獲得するかをまとめたい．本論文において，ロボットの知識はロボット自身の経験より得られる知覚情報を階層的なカテゴリに分類することで実現できると考える．図 2.20 に，提案するフレー

ムワークの全体像を示す．図中の各緑色のブロックにおいて，灰色のノードはロボットがセンシングできる情報を示しており，それらによって形成される物体概念（“O”）や動き概念（“M”）など多様な概念は上位のノードとなる．また，それらの概念の関係を表現する上位概念は図中の “I” で示している．

ロボットは，環境である人や物体などとインタラクションしながら知覚情報を得る．ロボットが得た知覚情報は，知識の学習に用いられる．知識の学習では，ロボットが観測する事物に含まれる共起性や推移性を手がかりとして利用する．共起性とは図 2.20 の緑色のブロックに示す，物体や場所などといった出来事の共起やある概念の形成に利用されるロボットによってセンシングされる知覚情報の共起などである．これを扱う計算モデルとして，階層的カテゴリ分類が可能なモデルである mMLDA を提案する．mMLDA については 3 章と 4 章で議論するが，図 2.20 に示すように，このモデルを用いることで概念や文法などの知識を獲得することが可能となる．

また，推移性とは続いて起こる出来事のことであり，例えば「お風呂を入って，テレビを見て，寝る」といった人の習慣がその一例である．これは，mMLDA によって表現される知識の時間的な順序（図 2.20 の青色の矢印）を考慮することを意味する．これを扱う計算モデルについては 5 章で述べ，人の習慣に含まれる行動の時間的な順序を考慮した行動文脈を獲得する手法について議論する．これによって例えば，人が行動しているときにロボットが先読みをして行動に必要な物体を届けるといったサービスを行うこともできる．

以上より，本論文ではロボットが知的に振る舞うためには，図 2.20 に示すような知識が必要であり，その知識は環境との相互作用により得た情報をもとに自律的に獲得しなければならないと考える．そのために，図中に表現される確率的知識をどのように獲得するかを次章以降に議論することとする．

第3章　人の動きと物体の関係による知識獲得

3.1　はじめに

前章では，ロボットのタスク実現に関する問題点を述べ，それを解決するための確率的知識表現の枠組みについて議論した．本章では，その議論に基づき確率的知識表現の核となる確率モデルを提案する．ここではまず，人の動きと物体の関係に関する知識獲得について述べる．さらに多様な概念への拡張や言語とのつながりについては，次章で議論する．

人間による事物の理解は，経験のカテゴリ分類によって形成される概念を通した予測に基づいていると考えることができる [10]．未知の環境でロボットが柔軟に動作するためにも，こうした概念形成は重要であり，近年そうした取り組みがなされている [55]．一方，人の動きを分節化し，それをカテゴリ分類することで概念化する研究も様々なされている．本章では，こういった動きと物の概念の関係を表すより高いレベルの概念を形成することを考える．これは，こうした概念間の関係の中にこそ知識が表現されていると考えられるためである．この目的のため本章では，階層的な概念の構造を形成可能な多層マルチモーダル LDA（mMLDA : multilayered Multimodal Latent Dirichlet Allocation）を提案する．

先行研究において，pLSA（probabilistic Latent Semantic Analysis）[56] や LDA（Latent Dirichlet Allocation）[57] を拡張したマルチモーダルカテゴリゼーションが提案され，複数のモダリティを用いることにより，より人間の感覚に近い物体カテゴリをロボットが教師なしで学習できることが示されている [18, 58]．ここで重要なのは，学習された物体カテゴリを基盤とした未観測情報の予測であり，これがロボットによる理解につながる [55]．また，こうした物体カテゴリが教師なし

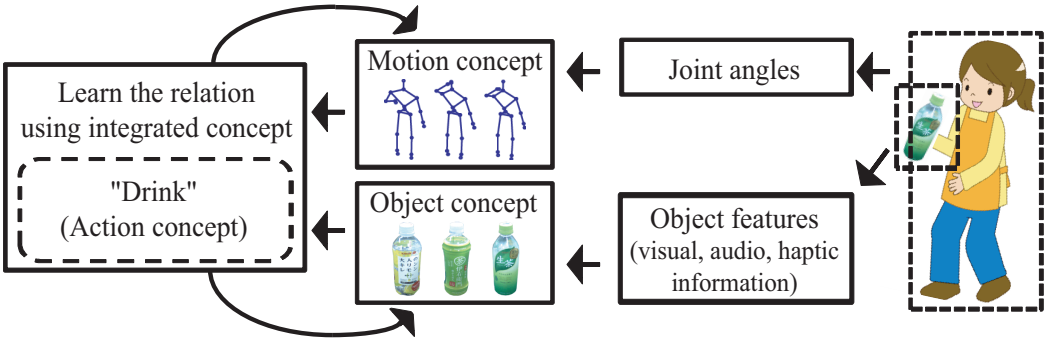

図 3.1: 統合概念形成の模式図

で学習されることが重要であり，学習された物体カテゴリを物体概念と呼ぶ [55].

しかし，ロボットが物体をより深く理解するためには，物体概念の学習だけでは不十分である．なぜなら，ほとんどの物体にはそれを使う人の動きが関連しており，物体とその物体に関連する人の動きを学習する必要があるためである．つまり，物体概念と人の動きの概念を学習すると同時に，それらの関係に埋め込まれた知識を獲得する必要があると言える．ここでは，これらの二つの概念が統合された上位の概念が動作概念であると考える．図 3.1 は，これを表現した模式図である．この図では,「飲み物」（物体概念）と「何かを口に運ぶ」（動き概念）という二つの下位概念が表現されている．さらにこれらの概念が統合されることで，より高いレベルの概念である「飲む」という概念（動作概念）が形成される．従って「飲む」という動作に関する知識は，飲み物という概念と口に運ぶという動き概念の関係に埋め込まれていることに他ならない．

またこの図において重要なのは，様々なレベルでの推論が可能であるという点である．上記の例においてロボットは，与えられたペットボトルの視覚的な情報から,「何か口に運ぶ」動きを想起することができる．逆に,「何か口に運ぶ」動きから,「ペットボトル」という物体を推論することも可能であり，これは動作に関する知識を利用した推論であると捉えることができる．さらに，上位概念の形成過程が下位概念の形成，つまりは物体や動きのカテゴリ分類に影響を及ぼすことは注目に値するであろう．例えば，全く異なるテクスチャをもちながらもボトルの形である物体は，飲み物とは別の物体カテゴリに分類される可能性があるが，こ

の物体が「何かを口に運ぶ」動きと共に使用される場合，統合概念である「飲む」が下位層の分類に影響することで,「ペットボトル」（物体概念）といった単一の物体概念を形成することに寄与する．一方で，物体が異なる動きに関係する場合，見た目の似た物体であっても異なるカテゴリに分類される可能性がある．

ここではこうした仕組みを実現するために，多層マルチモーダル LDA（multi-layered Multimodal LDA: mMLDA）を提案する．mMLDA は，下位層の物体概念と動き概念，および上位層でこれらを統合した統合概念で構成される．ロボットは学習プロセスにおいて，人の動きと使用される物体を観測する．物体概念は，ロボットが物体に関して取得したマルチモーダル情報，すなわち視覚，聴覚及び触覚情報をマルチモーダル LDA（MLDA）を用いることで形成する．同様に動き概念は，ロボットに搭載した KINECT から取得される人の関節角度情報に基づいて MLDA によって形成される．これら二つの MLDA は，上位の MLDA によって結合され，この上位層において下位概念間の関係性を表現するような上位概念，すなわち統合概念を形成することになる．ただし，全ての層の分類プロセスは相互に依存していることに注意が必要である．こうした相互依存的なモデルとは異なり，それぞれの概念を表現する複数の MLDA によって形成された概念を上位概念の入力とする簡易的な近似モデルを考えることも可能である．本論文ではこれを近似モデルと呼ぶことにする．一方，視覚・聴覚・触覚・動き情報を独立した下位概念として表現することも可能である．しかし，この場合，上位に形成される概念が物体概念となるのか，それとも物体と動きの関係を表す概念となるのかは明らかではなく，動き・物体・統合概念の 3 つを同時形成することはできない．また，先行研究により動きの情報から動き概念が形成できること，さらに視覚・聴覚・触覚情報から物体概念が形成できることが明らかとなっており [18, 58]，本章ではこれら二つの概念からそれらの関係を表現している統合概念を形成することを目的とする．

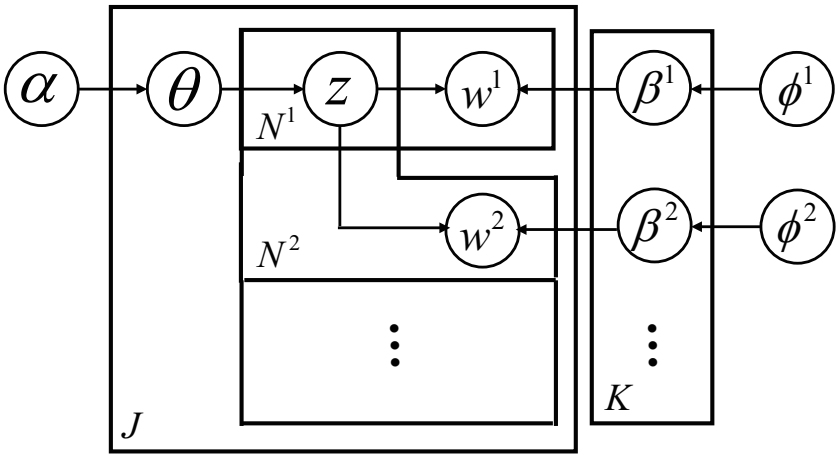

図 3.2: マルチモーダル LDA のグラフィカルモデル

3.2 マルチモーダル LDA

ここではまず，基本となる MLDA [58] について述べる．本論文で考える概念とは，ロボットが得られる情報をカテゴリ分類することで得られるカテゴリであり，MLDA はそのようなカテゴリを表現する生成モデルである．図 3.2 が MLDA のグラフィカルモデルであり，以下の手順で観測情報が生成されると考える．

1. カテゴリ z を生成する多項分布のパラメータ θ と，観測情報を生成する多項分布のパラメータ β_z^n $(n = 1, 2, \cdots)$ を，それぞれ α，ϕ^n をパラメータとするディリクレ事前分布から生成する

$$\theta \sim \mathrm{Dir}(\alpha) \tag{3.1}$$

$$\beta_z^n \sim \mathrm{Dir}(\phi^n) \tag{3.2}$$

2. 観測情報 n の i 番目のデータ w_i^n を，以下の処理を繰り返すことで生成する

   (a) カテゴリ z を θ をパラメータとする多項分布から生成する

$$z \sim \mathrm{Mult}(\theta) \tag{3.3}$$

(b) データ w_i^n を β_z^n をパラメータとする多項分布から生成する

$$w_i^n \sim \mathrm{Mult}(\beta_z^n) \tag{3.4}$$

ただし，図 3.2 における J はデータの総数であり，K はカテゴリ数を表す．また，N^* は各モダリティにおける情報の総数を表す．

ここでのカテゴリ分類の問題は，実際に観測した情報 $\boldsymbol{w}^n$ から，そのデータを生成するモデルのパラメータ θ および β_z^n を推定することに相当する．パラメータは，EM アルゴリズムや，ギブスサンプリングによって求めることが可能である．

また，学習した確率モデルを用いて，未知物体のカテゴリを推定することが可能である．未知物体のマルチモーダル情報 $\boldsymbol{w}'^1$，$\boldsymbol{w}'^2$，$\cdots$ が与えられた場合，そのカテゴリ $\hat{z}$ は $P(z|\boldsymbol{w}'^1, \boldsymbol{w}'^2, \cdots)$ を最大とするカテゴリを求めることで決定することができる．

$$\hat{z} = \underset{z}{\mathrm{argmax}} \int P(z|\theta)P(\theta|\boldsymbol{w}'^1, \boldsymbol{w}'^2, \cdots)d\theta \tag{3.5}$$

このように MLDA を用いてマルチモーダル情報を教師なしでカテゴリ分類することで，ロボットは自ら概念を形成することができる．これまで，物体から得られる視覚・聴覚・触覚情報を $\boldsymbol{w}^1$，$\boldsymbol{w}^2$，$\boldsymbol{w}^3$ と考えることで，MLDA により物体概念の形成が可能であるが示されている [58] ．また，概念は未観測情報の予測に利用することが可能であり，これがロボットによる事物の理解であると捉える．しかし，MLDA は物体など単一の概念しか表現することができず，また動きと物体など異なる概念間の関係性を捉えることができない．そこで次節において，複数の MLDA を階層的に結合した mMLDA に拡張することを考える．

3.3 概念の統合モデル

図 3.3 に，提案する多層 MLDA（mMLDA）のグラフィカルモデルを示す．図 3.3 において，z は統合概念を表すカテゴリであり，z^O，z^M はそれぞれ物体カテゴリと動きカテゴリであり，上位カテゴリ z により下位カテゴリである z^O と z^M

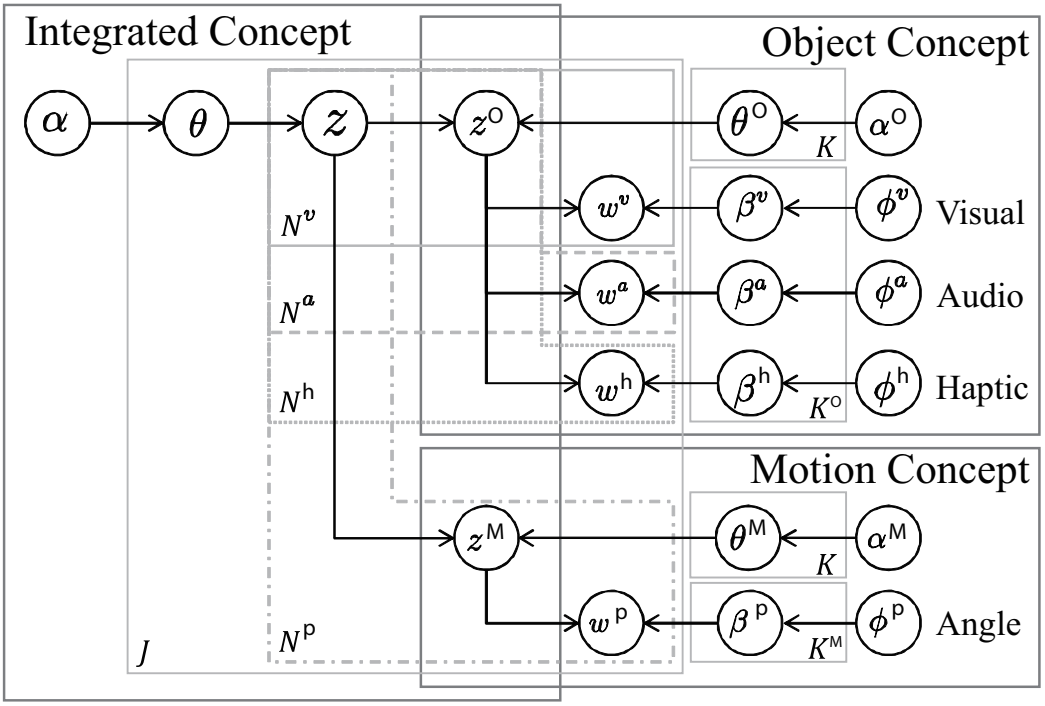

図 3.3: 多層マルチモーダル LDA のグラフィカルモデル

の関係を表現したモデルとなっている．また，w^v，w^a，w^h は，それぞれ物体から得られる視覚・聴覚・触覚情報であり，w^p は物体を扱っている際の人の動きの情報である．これらは，以下のように生成されることを仮定する．

1. 上位カテゴリ z を生成する多項分布のパラメータ θ と，概念 $C \in \{O, M\}$ を生成する多項分布のパラメータ θ_z^C と，各モダリティ $m \in \{v, a, h, p\}$ の情報を生成する多項分布のパラメータ β_{zC}^m を，それぞれ α，α^C，ϕ^m をパラメータとするディリクレ事前分布から生成する

$$\theta \sim \mathrm{Dir}(\alpha) \tag{3.6}$$

$$\theta_z^C \sim \mathrm{Dir}(\alpha^C) \tag{3.7}$$

$$\beta_z^m \sim \mathrm{Dir}(\phi^m) \tag{3.8}$$

2. 各概念の i 番目の情報 w_i^m を，以下の処理を繰り返すことで生成する

(a) 上位カテゴリ z を θ をパラメータとする多項分布から生成する

$$z \sim \mathrm{Mult}(\theta) \tag{3.9}$$

(b) カテゴリ z^C を，θ_z^C をパラメータとする多項分布から生成する

$$z^C \sim \mathrm{Mult}(\theta_z^C) \tag{3.10}$$

(c) カテゴリ z^C の情報 w_i^m を $\beta_{z^C}^m$ をパラメータとする多項分布から生成する

$$w_i^m \sim \mathrm{Mult}(\beta_{z^C}^m) \tag{3.11}$$

ただし，図 3.3 における J はデータの総数であり，K と K^C はそれぞれ上位と下位概念のカテゴリ数を表す．また，N^m は各モダリティにおける情報の総数を表す．この生成過程では，上位カテゴリ z が決まると，それと対応した θ_z^O と θ_z^M から物体カテゴリと動きカテゴリが生成されており，z によって z^O と z^M の関係が表現されている．

3.3.1 物体概念

物体概念は，ロボットが実際に取得したマルチモーダル情報をカテゴリ分類することにより形成する．図 3.3 の物体概念部分だけを見ると，図 3.2 と等価なモデルであり，視覚・聴覚・触覚情報 w^v, w^a, w^h がその類似性により分類され，物体カテゴリ z^O を形成することができる．マルチモーダル情報は，図 3.4 に示した情報取得が可能なアームロボットにより取得する．

視覚情報　図 3.4 に示したアームロボットを用いて物体を様々な角度から観測し，物体毎に画像 7 枚を取得する．特徴量として 128 次元の DSIFT [44] を用い，これにより 1 枚の画像から多数の特徴ベクトルを得る．これらの特徴ベクトルを学習画像とは関係のない背景画像から計算した 500 の代表ベクトルを用いてベクトル量子化し，500 次元のヒストグラムとする．

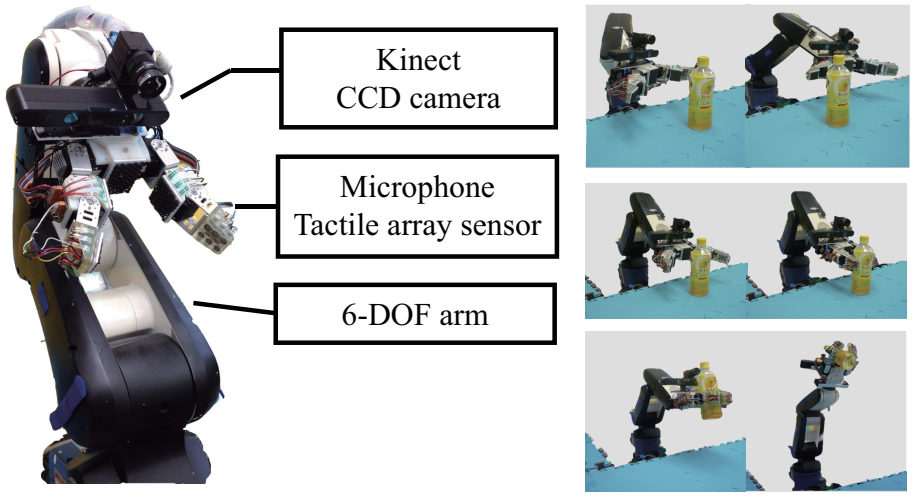

図 3.4: ロボットとマルチモーダル情報取得：(a) アームロボット (b) 視覚情報（上），触覚情報（中），聴覚情報（下）

触覚情報　触覚情報には，各物体を数回握り，32 個のセンサから構成された触覚アレイセンサにより取得した時系列データを用いる．取得したデータは曲線近似を行い，そのパラメータを各センサの特徴ベクトルとして扱う [59]．さらに K 平均法により予め計算した 15 の代表ベクトルを用いてベクトル量子化を行い，最終的に得られる 15 次元ヒストグラムを触覚情報として利用する．

聴覚情報　各物体を振った際に取得した音信号を 0.2[s] 毎のフレームに分割し，フレーム毎の特徴ベクトルに変換する．特徴量としては，音声認識で最もよく使用されている MFCC を用いることとし，これにより各フレームは 13 次元の特徴ベクトルとなる．この特徴ベクトルを，予め計算した 50 の代表ベクトルを用いてベクトル量子化し，50 次元ヒストグラムとして扱う．

3.3.2 動き概念

前述の物体概念と同様に，図 3.3 における w^p は人が物体に対して行う動きの情報を表す．また，図 3.3 における動き概念に相当する部分だけを見ると，図 3.2 と等価なモデルであり，w^p がその類似性により分類され，動きカテゴリ z^M を形成す

ることになる．動き情報は，人が物体を使用する動作中の関節角度を KINECT により取得し，動き毎にデータを 5 回記録し計算する．取得した関節角は 11 箇所であり，一つの物体の使用開始から使用終了までを連続して取得した．ただし，動きの情報を分類するためには，関節角の時系列データを単位動作毎に分節化する必要がある．時系列データの分節化については，階層ディリクレ過程隠れマルコフモデル（Hierarchical Dirichlet Processes Hidden Markov Model: HDP-HMM）[60] などを用いることが可能であるが，ここでは分節化の問題を簡単にするために，使用している物体の同一性を手掛かりに動きの情報を分節化できると仮定する．つまり，ある物体を使い続けている間は一つの動作であると考え，物体の使い始めから使い終わりまでを一つの分節と考える．

11 箇所から得られる関節角を，11 次元の特徴ベクトルと考えることで，一つの動作から 11 次元の特徴ベクトルを複数得ることができる．それらを予め計算した 70 の代表ベクトルによりベクトル量子化することで得られる 70 次元のヒストグラムを動き情報として用いる．これは動きの位相情報を捨て去った Bag of Features（BoF）表現である．このような特徴量を動作認識に使用することは，文献 [61] において提案されており，その有効性が示されている．

3.3.3 統合モデル

提案モデルでは，物体概念と動き概念が，それぞれ MLDA で表現されており，さらにその上位で二つの概念の関係を統合概念として表現する二層構造のモデルとなっている．物体概念，動き概念を独立した MLDA として学習し，その学習結果を統合概念として学習することも可能であるが，そのような方法では物体概念と動き概念が相互に影響をあたえることができない．そこで，ここでは各概念を表す隠れ変数 z，z^O，z^M を同時に学習する．学習にはギブスサンプリングを用い，各概念を表すカテゴリ z，z^O，z^M を，観測データ w^v，w^a，w^h，w^p からサンプリングすることで学習する．サンプリングには，θ，θ^O，θ^M，β^m を周辺化した以下

の事後分布を用いる．

$$
\begin{aligned}
&P(z_{jmi}, z^O_{jmi}, z^M_{jmi} | \boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^O_{-jmi}, \boldsymbol{Z}^M_{-jmi}, \boldsymbol{W}^v, \boldsymbol{W}^a, \boldsymbol{W}^h, \boldsymbol{W}^p) \\
&= P(z_{jmi}|\boldsymbol{Z}_{-jmi}) P(z^O_{jmi}|z_{jmi}, \boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^O_{-jmi}) \\
&\quad \times P(z^M_{jmi}|z_{jmi}, \boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^M_{-jmi}) \\
&\quad \times P(w^v_{ji}|z^O_{jmi}, \boldsymbol{Z}^O_{-jmi}, \boldsymbol{W}^v_{-ji}) \\
&\quad \times P(w^a_{ji}|z^O_{jmi}, \boldsymbol{Z}^O_{-jmi}, \boldsymbol{W}^a_{-ji}) \\
&\quad \times P(w^h_{ji}|z^O_{jmi}, \boldsymbol{Z}^O_{-jmi}, \boldsymbol{W}^h_{-ji}) \\
&\quad \times P(w^p_{ji}|z^M_{jmi}, \boldsymbol{Z}^M_{-jmi}, \boldsymbol{W}^p_{-ji})
\end{aligned} \tag{3.12}
$$

右辺のそれぞれの確率分布は次のようになる．

$$
P(z_{jmi} = k|\boldsymbol{Z}_{-jmi}) = \frac{\alpha + N^{-jmi}_{j,z=k}}{K\alpha + N_j{}^{-jmi}} \tag{3.13}
$$

$$
\begin{aligned}
&P(z^C_{jmi} = l | z_{jmi} = k, \boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}) \\
&\qquad = \frac{\alpha^C + N^{-jmi}_{z=k,z^C=l}}{K^C\alpha^C + N^{-jmi}_{z=k}}
\end{aligned} \tag{3.14}
$$

$$
\begin{aligned}
&P(w^m_{ji} = x | z^C_{jmi} = k, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{W}^m_{-ji}) \\
&\qquad = \frac{\phi^m + N^{-jmi}_{z^C=k,w^m=x,m}}{W^m\phi^m + N^{-jmi}_{z^C=k,m}}
\end{aligned} \tag{3.15}
$$

ただし，$\boldsymbol{Z}$，$\boldsymbol{Z}^O$，$\boldsymbol{Z}^M$ は，それぞれ全物体の全情報に割り当てられた上位カテゴリ，物体カテゴリ，動きカテゴリの集合を表し，$\boldsymbol{W}^m$ はモダリティ m の全物体の情報の集合である．$N_{j,z=k}$ は物体 j の全モダリティの上位カテゴリ z に k が割り当てられた回数であり，$N_{z^C=k,w^m=x,m}$ はモダリティ m の特徴量 w^m に x が，下位カテゴリ z^C に k が割り当てられた回数である．また，$N_{z=k,z^C=l}$ は上位カテゴリ $z=k$ と下位カテゴリ $z^C=l$ の共起した回数を表しており，K，K^C，W^m はそれぞれ上位カテゴリのカテゴリ数，概念 C のカテゴリ数，モダリティ m の情報の次元数である．負の添字はその情報を除外することを表し，$-jmi$ は j 番目の物体のモダリティ m の i 番目の情報を除外することを表している．

Algorithm 1 Multilayered MLDA (bottom layer)

for all i, j, C, m **do**
  $u \leftarrow$ draw from Uniform [0,1]
  for $k \leftarrow 1$ to K^C **do**
    $P[k] \leftarrow P[k-1] + P(z^C_{jmi} = k | w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi})$
  end for
  for $k \leftarrow 1$ to K^C **do**
    if $u < P[k]/P[K^C]$ **then**
      $z^C_{jmi} = k$, break
    end if
  end for
end for

Algorithm 2 Multilayered MLDA (whole layer)

for all i, j, C, m **do**
  for $k \leftarrow 1$ to K **do**
    $P[k] \leftarrow P[k-1] + P(z_{jmi} = k | w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi})$
  end for
  $u \leftarrow$ draw from Uniform [0,1]
  for $k \leftarrow 1$ to K **do**
    if $u < P[k]/P[K]$ **then**
      $z_{jmi} = k$, break
    end if
  end for
  for $k \leftarrow 1$ to K^C **do**
    $P[k] \leftarrow P[k-1] + P(z^C_{jmi} = k | w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi})$
  end for
  $u \leftarrow$ draw from Uniform [0,1]
  for $k \leftarrow 1$ to K^C **do**
    if $u < P[k]/P[K^C]$ **then**
      $z^C_{jmi} = k$, break
    end if
  end for
end for

モデルの学習は，隠れ変数である z，z^O，z^M を，収束するまで事後分布からサンプリングすることによって実現できる．しかし，隠れ変数が 3 つあり，複雑なモデルであるため，全てのパラメータを同時に求めると局所解に陥りやすいといった問題がある．そこで，図 3.3 の右側に示す下位カテゴリ z^C を個々の独立した MLDA として学習し，下位概念のパラメータ β^m（式（3.15））を先に決定する．

次に，式（3.15）を固定し，上位カテゴリ z，下位カテゴリ z^O，z^M をサンプリングする．

$$\begin{aligned} z^C_{jmi} &\sim P(z^C_{jmi}|w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi}) \\ &\propto \sum_z P(z|\boldsymbol{Z}_{-jmi}) P(z^C_{jmi}|\boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}, z) \\ &\quad \times P(w^m_{ji}|\boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, z^C_{jmi}) \end{aligned} \tag{3.16}$$

$$\begin{aligned} z_{jmi} &\sim P(z_{jmi}|w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi}) \\ &\propto \sum_{z^C} P(z_{jmi}|\boldsymbol{Z}_{-jmi}) P(z^C|\boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}, z_{jmi}) \\ &\quad \times P(w^m_{ji}|\boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, z^C) \end{aligned} \tag{3.17}$$

このとき，下位カテゴリ z^C が上位概念の影響を受けて更新されることに注意が必要である．Algorithm 1 と Algorithm 2 がそれぞれ，下位概念のパラメータの決定と，モデル全体の学習アルゴリズムである．以上のようなサンプリングを繰り返すことで，N_* がある値へと収束する．K を上位カテゴリのカテゴリ数とするとき，最終的なパラメータの推定値 $\hat{\beta}^m_{w^m z^C}$，$\hat{\theta^C}_{zz^C}$，$\hat{\theta}_{jz}$ は以下のようになる．

$$\hat{\beta}^m_{w^m z^C} = \frac{N_{z^C w^m m} + \phi^m}{N_{z^C m} + W^m \phi^m}, \tag{3.18}$$

$$\hat{\theta^C}_{zz^C} = \frac{N_{zz^C m} + \alpha^C}{N_{zm} + K\alpha^C}, \tag{3.19}$$

$$\hat{\theta}_{jz} = \frac{N_{jz} + \alpha}{N_j + K\alpha}, \tag{3.20}$$

ただし，W^m はモダリティ m の次元数を表し，$N_{z^C w^m m}$ はモダリティ m の w^m に下

位カテゴリ z^C が割り当てられた回数を表す．

学習したモデルを用いることで，物体や動作の認識だけでなく，概念間の予測も可能となる．例えば，物体の視覚 $\boldsymbol{w}^v$・聴覚 $\boldsymbol{w}^a$・触覚 $\boldsymbol{w}^h$ 情報が得られた際に，以下の式を用いて，物体カテゴリ $\hat{z}^O$，その物体に関係する上位カテゴリ $\hat{z}$ と動きカテゴリ $\hat{z}^M$ を予測することができる．

$$\hat{z}^O = \underset{z^O}{\operatorname{argmax}} \sum_z \sum_{z^M} \hat{P}(z, z^O, z^M | \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h) \tag{3.21}$$

$$\hat{z}^M = \underset{z^M}{\operatorname{argmax}} \sum_z \sum_{z^O} \hat{P}(z, z^O, z^M | \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h) \tag{3.22}$$

$$\hat{z} = \underset{z}{\operatorname{argmax}} \sum_{z^O} \sum_{z^M} \hat{P}(z, z^O, z^M | \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h) \tag{3.23}$$

ただし，$\hat{P}(z, z^O, z^M | \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h)$ は以下のように計算することができる．

$$\hat{P}(z, z^O, z^M | \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h) = P(z)P(z^M, z^O | z)P(z^O | \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h) \tag{3.24}$$

また，同様に，動きの情報 $\boldsymbol{w}^p$ から各概念のカテゴリを予測するには，式（3.24）の代わりに次式を用いることで可能となる．

$$\hat{P}(z, z^O, z^M | \boldsymbol{w}^p) = P(z)P(z^O, z^M | z)P(z^M | \boldsymbol{w}^p) \tag{3.25}$$

3.3.4 近似モデル

前述のように，物体概念と動き概念は，統合概念を無視すれば，独立した MLDA と等価なモデルと考えることができる．さらに，視覚・聴覚・触覚・動き情報 w^m を無視することで，統合概念は z^O と z^M を生成する MLDA と等価なモデルと見なすことができる．すなわち各概念を独立した MLDA として学習し，形成された概念を上位概念の入力とすることで，簡易的に物体概念，動き概念を統合することができる．図 3.5 が mMLDA を分解し，独立した 3 つの MLDA として考えた場合のグラフィカルモデルである．このモデルでは，物体カテゴリ z^O と動きカテゴ

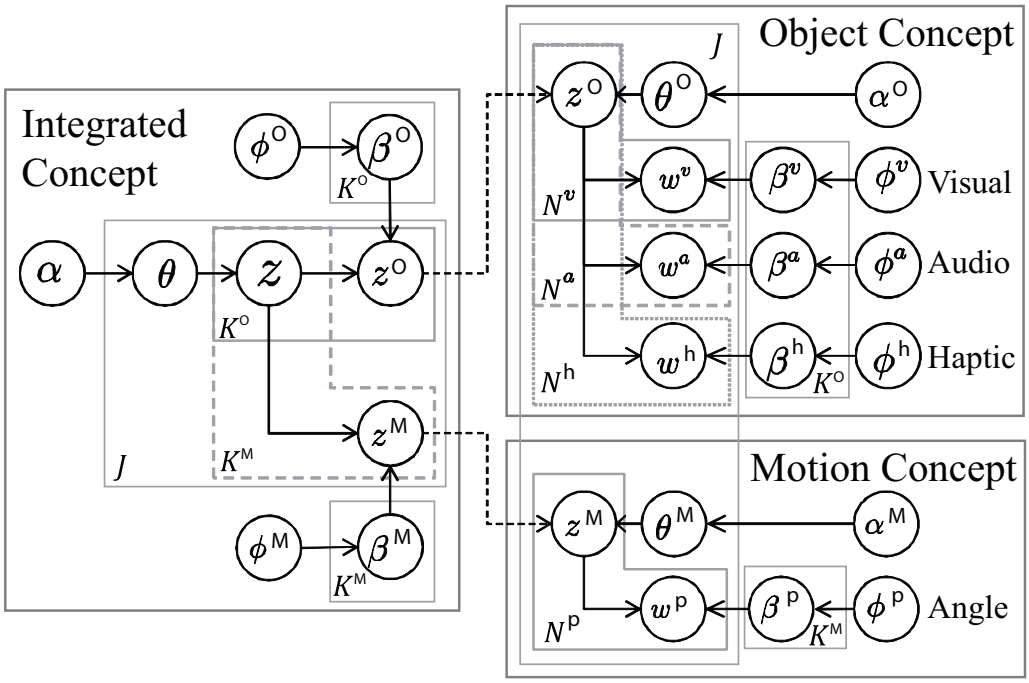

図 3.5: 統合概念の近似モデル

リ z^M を独立した MLDA で学習後，上位カテゴリ z をもう一つの独立な MLDA で学習すればよい．

近似モデルの学習ではまず，下位層を独立した MLDA として学習した後に，z^O と z^M をそれぞれ多項分布 $P\left(z^O \mid \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h\right)$ と $P\left(z^M \mid \boldsymbol{w}^p\right)$ からサンプリングする．近似モデルの上位層に相当する MLDA では，生成された z^O と z^M をそれぞれ図 3.2 に示した w^1 と w^2 として考えることで学習することができる．従って，物体概念と動き概念の関係性は，モデルにおける隠れ変数 z によって学習され，この z が両概念の統合的な概念（動作概念）を表現するカテゴリとなる．ただし，z は固定された z^O と z^M の関係性を表現するだけであり，逆に z^O と z^M に影響を与えることはない．

近似モデルにおいても，学習したモデルを用いて未観測情報を予測することが可能である．例えば，物体情報 $\boldsymbol{w}^v$，$\boldsymbol{w}^a$，$\boldsymbol{w}^h$ から，確率の高い動きカテゴリを次のように予測することができる．

1. 物体情報から物体カテゴリをサンプリングする

$$\hat{z}^O \sim P\left(z^O \mid \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h\right) \tag{3.26}$$

2. 次式によりサンプリングされた物体カテゴリ $\hat{\boldsymbol{z}}^O$ から，動作カテゴリが発生する確率を計算する

$$P(\hat{z}^M|\hat{\boldsymbol{z}}^O) = \int \sum_z P(\hat{z}^M|z)P(z|\theta)P(\theta|\hat{\boldsymbol{z}}^O)d\theta \tag{3.27}$$

同様に，動き情報から最も関係する物体を予測することも可能である．後に示す実験の結果からも分かるように，mMLDA と近似モデルの定性的差異は明らかである．近似モデルは単純かつ容易に実装できるところがメリットではあるが，明らかな欠点を持つ．それは，上位概念が下位概念に一切影響を与えないことである．各概念を独立に学習することになるため，下位概念での分類誤りがそのまま上位概念の学習に影響を及ぼし，モデル全体の精度を下げることにつながる．一方，mMLDA は，各概念が同時に形成されるために，下位層での分類が相互に影響を及ぼし合い，モデル全体として分類や予測の精度を向上させることができると期待できる．以降，提案する mMLDA の有効性を評価するために，近似モデルとの比較を行う．

3.4 実験

提案した mMLDA を評価するために実験を行った．図 3.6 に，実験に使用した物体を示す．これらの物体を使用する動作を行い，実験のためのデータを取得した．物体と動きの組み合わせを，表 3.1 に示す．

物体概念は，ロボットが図 3.6 の各物体を観測することで取得するマルチモーダル情報を基に形成する．また動き概念は，人が物体を使用している際の動きを，KINECT を用いて取得した関節角情報を基に形成する．人が使用している物体は，ロボットが視覚的にトラッキングできることを仮定している．つまり，物体に関

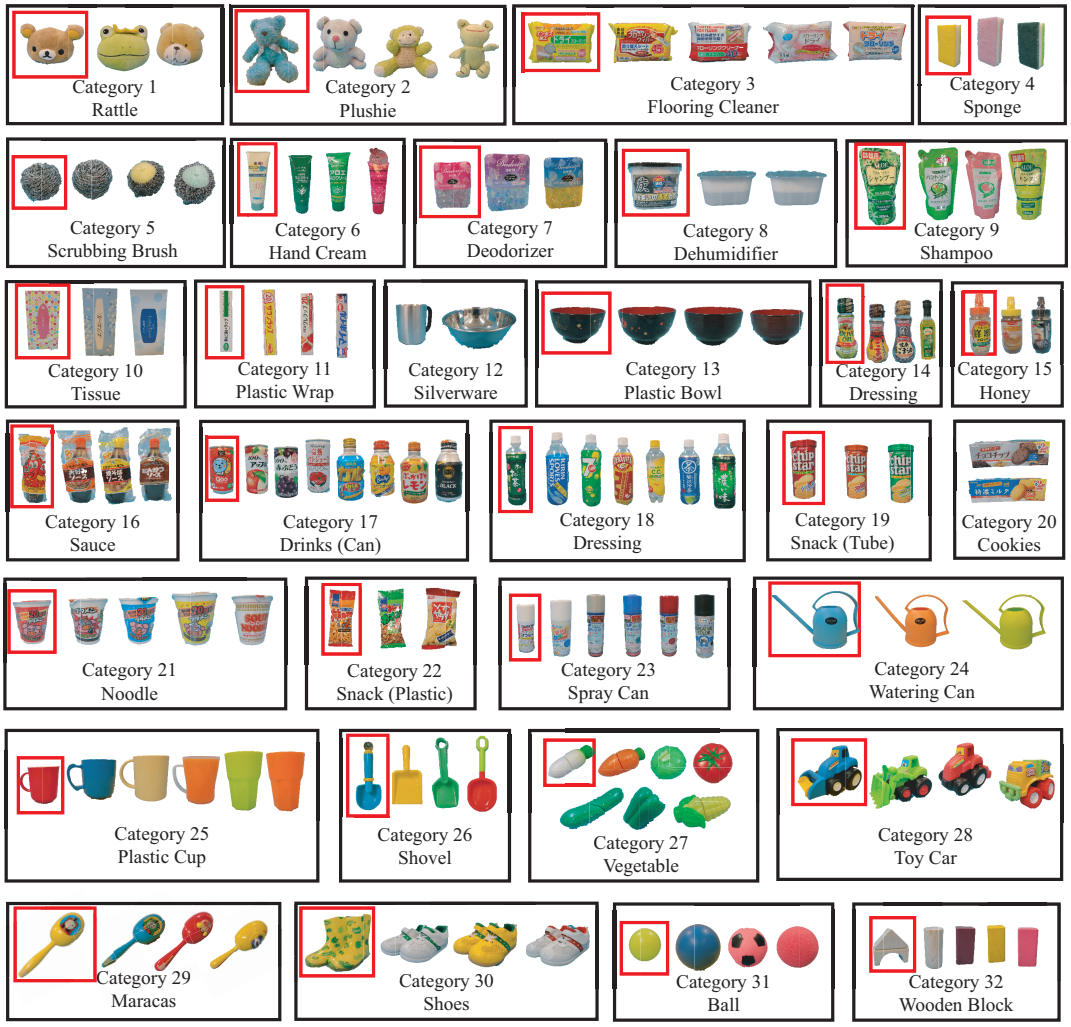

図 3.6: 実験で使用した物体（各カテゴリ内の枠は認識用の物体）

するマルチモーダル情報とその物体を使用した際の関節角情報はモデルを学習するために同時に与えられることになる．図 3.7 に，取得した動きの画像とデータの一部を示す．

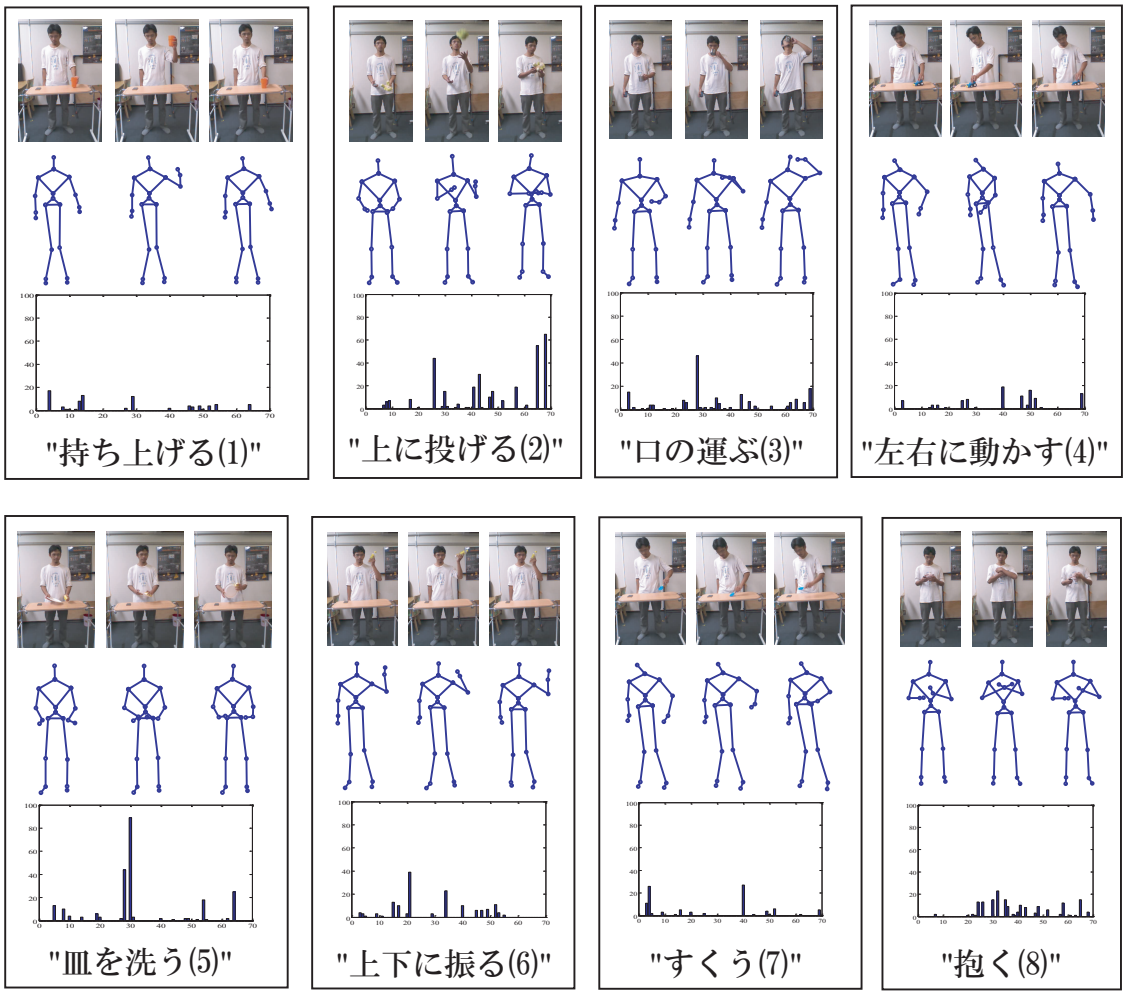

図 3.7: 各動きから取得した情報の例：（上から下まで）実際の動き，KINECT から取得した情報，70 次元のヒストグラム（括弧内の数字はカテゴリ番号）

3.4.1 カテゴリ数決定

LDA ではカテゴリ数を予め与えなければならず，このカテゴリ数の決定は LDA における重要な問題である．本論文で提案する mMLDA は LDA を拡張したモデルであるため，同様の問題が存在する．予備実験などを通して経験的に決定することも可能ではあるが，特に上位層の分類に対する正解を決めることは人手であっても困難であるため，ここでは自動的にカテゴリ数を決めることを考える．

表 3.1: 物体に対して行った動き（括弧内の数字はカテゴリ番号）

動き	物体	動き	物体
持ち上げる（1）	茶碗（13）	食べる（9）	野菜（玩具）（27）
	飲み物（缶）（17）		スナック（19）
	カップヌードル（21）		カップヌードル（21）
	スプレー缶（23）	積み重ねる（10）	積み木（32）
	プラスチックカップ（25）	置く（11）	消臭剤（7）
上に投げる（2）	ぬいぐるみ（2）		湿気取り（8）
	マラカス（29）		プラスチックカップ（25）
	ボール（31）		積み木（32）
片手で口に運ぶ（3）	金属のカップ（12）	手に塗る（12）	ハンドクリーム（6）
	飲み物（缶）（17）	取り出す（13）	フローリングワイパー（3）
	ペットボトル（18）		ティッシュ箱（10）
	プラスチックカップ（25）		クッキー（20）
左右に動かす（4）	車（玩具）（28）	ナイフで切る（14）	野菜（玩具）（27）
	フローリング ワイパー（3）	中身をかける（15）	ドレッシング（14）
			蜂蜜（15）
皿を洗う（5）	スポンジ（4）		ソース（16）
	たわし（5）	中身を注ぐ（16）	シャンプー（9）
上下に振る（6）	ガラガラ（1）		飲み物（缶）（17）
	ドレッシング（14）		ペットボトル（18）
	ソース（16）		じょうろ（24）
	飲み物（缶）（17）	包む（17）	ラップ（11）
	ペットボトル（18）	塗る（18）	スプレー缶（23）
	スプレー缶（23）	履く（19）	靴（30）
	マラカス（29）	開ける（20）	スナック（22）
すくう（7）	ショベル（26）	両手で口に運ぶ（21）	金属のカップ（12）
抱く（8）	ぬいぐるみ（2）		茶碗（13）

下位層については，ノンパラメトリックベイズ手法であるマルチモーダル階層ディリクレ過程（Multimodal Hierarchical Dirichlet Process: MHDP）[62] による決定手法がそのまま利用できる．実際に MHDP によって下位層のカテゴリ数を推定したところ，物体と動きのカテゴリ数はそれぞれ 32 と 21 であった．この結果を用いて以降の実験を行うこととし，更に上位層のカテゴリ数を推定するために用いる．

上位層のカテゴリ数は，MHDP を直接適用して推定することができない．そこで，近似モデルの上位 MLDA に MHDP を適用することでカテゴリ数を推定することとする．MHDP ではサンプリングにより学習を行なっているため，初期値によって推定されるカテゴリ数が変わってしまう．そこで，MHDP を用いた分類を 100 回行い，100 個のモデルを学習した．図 3.8 が 100 個のモデルのカテゴリ数のヒストグラムであり，横軸と縦軸はそれぞれ，推定した上位カテゴリ数とその頻度を示している．すなわち，このグラフはカテゴリ数の発生確率と考えることが

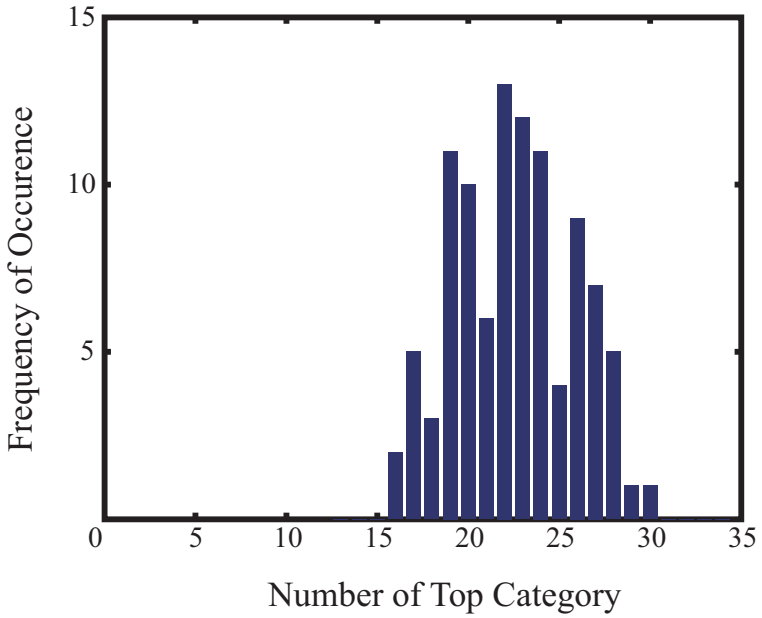

図 3.8: MHDP を用いたカテゴリ数の発生頻度

でき，カテゴリ数 22 が最も高い確率で発生していることが分かる．

以上の結果から，上位カテゴリ数を 22，物体と動きのカテゴリ数はそれぞれ 32 と 21 として，mMLDA と近似モデルによって概念形成を行い，それらを評価する．

3.4.2 物体概念

提案モデルと近似モデルによって形成された物体概念を評価した．物体概念の形成結果は図 3.9 であり，縦軸が物体の番号，横軸がモデルによって分類されたカテゴリを表している．図 3.9（a）が人手による分類であり，これを正解として各手法の分類結果を評価した．図 3.9（b）が提案手法（mMLDA）による分類結果であり，図 3.9（c）が近似モデルによる分類結果である．これらの分類結果から，図 3.9（a）を正解として，次式により分類精度を計算した．

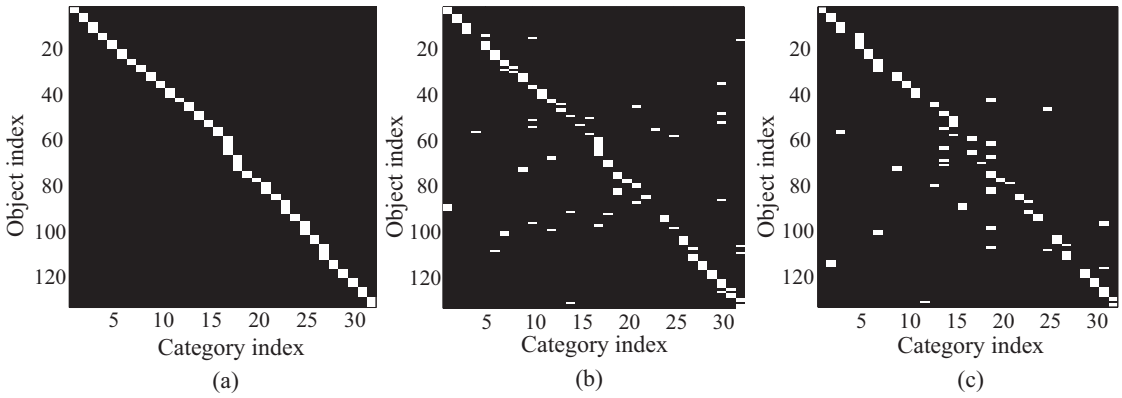

図 3.9: 物体の分類結果：(a) 正解, (b) mMLDA, (c) 近似モデル

$$Acc = \frac{100}{J}\sum_{j}^{J}\delta(c_{\text{correct}}(j), c_{\text{result}}(j)) \tag{3.28}$$

ただし，J はデータ数，$c_{\text{correct}}(j)$，$c_{\text{result}}(j)$ はそれぞれ j 番目のデータの正解のカテゴリと，実際に分類されたカテゴリの ID である．$\delta(a,b)$ は，$a=b$ のとき 1，さもなくば 0 となる関数である．分類精度を計算した結果，mMLDA では 71.21%，近似モデルでは 65.15%となり，提案モデルである mMLDA の方がより正解に近い分類ができている．mMLDA の分類では,「飲み物（缶）(17)」は一つのカテゴリに分類することができたのに対して，近似モデルでは，この物体を 3 つのカテゴリに分類してしまっている．同じ「飲み物（缶）(17)」でも，異なる柄や形を持つため，近似モデルでは異なるカテゴリに分類されてしまったのに対して，mMLDA では「飲み物（缶）(17)」と関係する動きも考慮して分類を行うため，正しく一つのカテゴリに分類することができたと考えられる．

3.4.3 動き概念

次に，提案モデルと近似モデルによって分類された動き概念を評価した．図 3.10 が分類結果であり，縦軸が実際の動きの番号，横軸が分類されたカテゴリ番号である．図 3.10 (a) が人手による分類であり，物体と同様，この分類を正解として，

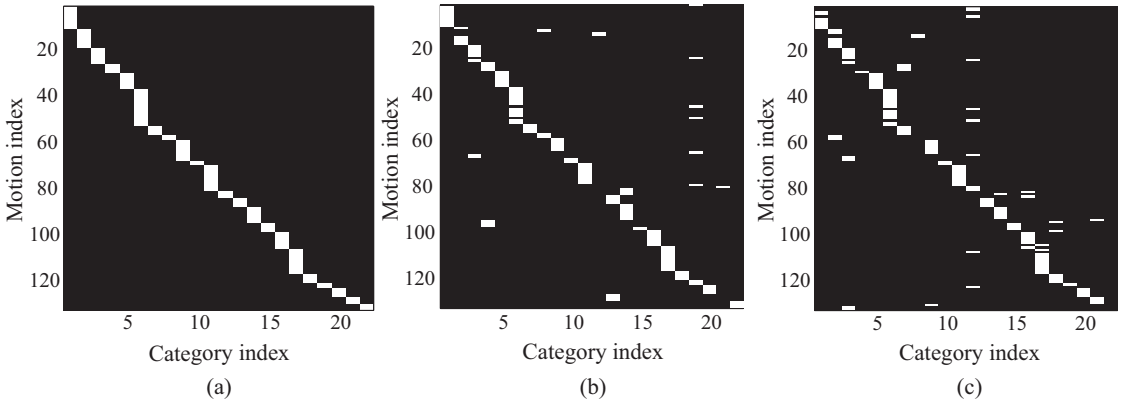

図 3.10: 動きの分類結果：(a) 正解, (b) mMLDA, (c) 近似モデル

各手法の分類を評価した．図 3.10（b）が mMLDA による分類結果，図 3.10（c）が近似モデルによる分類結果である．正解の分類（図 3.10（a））と比較すると，mMLDA（図 3.10（b））の分類精度は 84.09%となり，近似モデル（図 3.10（c））の分類精度は 77.27%となった．

mMLDA と近似モデルによる動き概念の形成結果の差異は,「左右に動かす（4)」と「すくう（7)」の分類結果で見ることができる．mMLDA の分類結果では，この二つの動きを二つのカテゴリに分類することができた．一方，近似モデルではこの二つの動きを一つのカテゴリに分類してしまっている．二つの動きは似通っているが，扱う物体が異なるため，mMLDA では二つのカテゴリに分類することができたと考えられる．このように，mMLDA は近似モデルに比べて，物体と動きがそれぞれ影響し合うため，より正解に近い分類が可能となる．

3.4.4 統合概念

次に上位層で形成された統合概念の評価を行った．まず，形成された統合概念の結果について述べる．mMLDA の上位層では物体と動きの関係性を表すカテゴリが形成されており，その中には人にとって意味のあるカテゴリも形成されている．表 3.2 が実際に形成された物体概念と動き概念が組み合わさり形成された統合概念である．例えば，統合概念 10 では，動きの「片手で口に運ぶ（3)」と物体の

表 3.2: mMLDA を用いた統合概念の形成結果（括弧内の数字はカテゴリ番号）

No	動き	物体
1	置く（11）	消臭剤（7）
		湿気取り（8）
		プラスチックカップ（25）
2	ナイフで切る（14）	野菜（玩具）（27）
3	包む（17）	ラップ（11）
4	抱く（8）	ぬいぐるみ（2）
5	左右に動かす（4）	車（玩具）（28）
6	左右に動かす（4）	フローリングワイパー（3）
7	持ち上げる（1）	茶碗（13）
		飲み物（缶）（17）
		カップヌードル（21）
		スプレー缶（23）
		プラスチックカップ（25）
8	上下に振る（6）	ガラガラ（1）
		ドレッシング（14）
		ソース（16）
		飲み物（缶）（17）
		ペットボトル（18）
		マラカス（29）
9	中身を注ぐ（16）	シャンプー（9）
		飲み物（缶）（17）
		ペットボトル（18）
		じょうろ（24）
10	片手で口に運ぶ（3）	金属のカップ（12）
		飲み物（缶）（17）
		ペットボトル（18）
		プラスチックカップ（25）
11	積み重ねる（10）	積み木（32）
	置く（11）	
12	上に投げる（2）	ぬいぐるみ（2）
		マラカス（29）
		ボール（31）
13	手に塗る（12）	ハンドクリーム（7）
14	皿を洗う（5）	スポンジ（4）
		たわし（5）
15	両手で口に運ぶ（21）	金属のカップ（12）
		茶碗（13）
	食べる（9）	野菜（玩具）（27）
		スナック（19）
		カップヌードル（21）
16	塗る（18）	スプレー缶（23）
17	履く（19）	靴（30）
18	取り出す（13）	フローリングワイパー（3）
		ティッシュ箱（10）
		クッキー（20）
19	開ける（20）	スナック（22）
20	中身をかける（15）	ドレッシング（14）
		蜂蜜（15）
		ソース（16）
21	すくう（7）	ショベル（26）
22	上下に振る（6）	スプレー缶（23）

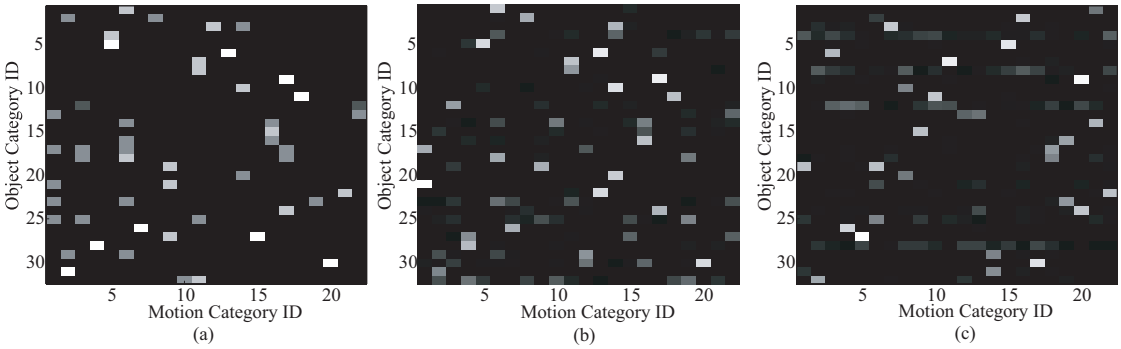

図 3.11: 物体カテゴリと動きカテゴリの共起確率 : (a) 正解, (b) mMLDA, (c) 近似モデル

「飲み物（缶）(17)」や,「ペットボトル（18)」,「プラスチックカップ（25)」,「金属のカップ（12)」などが一つのカテゴリに分類された．これは,「何かを飲む」という概念が形成されたことを意味する．他にも，統合概念 15 では「両手で口に運ぶ（21)」と「食べる（9)」といった動きと,「茶碗（13)」や「スナック（19)」などが組み合わさった概念が形成され，これは「何かを食べる」という概念であると言える．また，統合概念 9 では,「中身を注ぐ（16)」と,「ペットボトル（18)」や「じょうろ（24)」などが組み合わさった概念が形成され，これは「何かを注ぐ」といった概念であると言える．さらに，統合概念 5 と 6 では,「左右に動かす（4)」が，物体によって異なる上位カテゴリに分類された．一つは「車（玩具）(28)」と関係し，もう一つは「フローリングワイパー（3)」と関係する上位カテゴリであり，これらはそれぞれ「車の玩具を走らせる」という概念と,「フローリングワイパーで掃除をする」といった概念であると言える．このように同じ動きに対しても，異なる物体によって，意味の異なる統合概念が形成されている．

以上のように，定性的には意味のある統合概念が形成できたと言えるが，統合概念は正解を定義することが難しいため，定量的に mMLDA と近似モデルを比較することができない．そこで，ここでは物体と動きの関係性を正確に表現できているかどうかで評価する．物体カテゴリ z^O と動きカテゴリ z^M の関係性は，その同時確率 $P(z^O, z^M)$ で表現することができる．正解となる同時確率 $\hat{P}(z^O, z^M)$ は，

表 3.1 に示した各物体と動きの関係の学習サンプル数から，次式を用いて求めた．

$$\hat{P}(z^O, z^M) = \frac{N_{z^O,z^M}}{N} \tag{3.29}$$

ただし，N_{z^O,z^M} は，物体カテゴリ z^O と動きカテゴリ z^M の共起したデータ数であり，表 3.1 から求めることができる．また，N はデータの総数である．図 3.11（a）が，色の濃淡で正解の同時確率を表現したグラフである．縦軸と横軸は，それぞれ物体と動きのカテゴリ番号を表す．また，mMLDA と近似モデルで学習された同時確率は $P(z^O, z^M)$ は，次のように計算可能である．

$$P(z^O, z^M) = \sum_z P(z^O|z)P(z^M|z)P(z|\alpha) \tag{3.30}$$

図 3.11（b）と（c）が，それぞれ mMLDA と近似モデルによって学習された物体カテゴリと動きカテゴリの同時確率である．ここでは学習された同時確率 $P(z^O, z^M)$ がどれだけ正解 $\hat{P}(z^O, z^M)$ に近いかで評価し，その評価基準として次式で定義される Kullback-Leibler（KL）ダイバージェンスを用いた．

$$D_{KL}\left(P(z^O, z^M) \| \hat{P}(z^O, z^M)\right) = \sum_{z^O}\sum_{z^M} P(z^O, z^M) \log \frac{P(z^O, z^M)}{\hat{P}(z^O, z^M)} \tag{3.31}$$

KL ダイバージェンスは，二つの確率分布に対してそれらの間の差異を測るものであり，各モデルと正解基準との違いを表している．近似モデルの結果と mMLDA の結果の正解との KL ダイバージェーンスを求めたところ，それぞれ 6.26 と 4.17 となり，mMLDA の学習結果が正解に近いという結果となった．すなわち，mMLDA の方が近似モデルに比べ，より正確に物体と動きの関係，つまりは動作に関する知識を捉えている．

また実験では，上位カテゴリ数はノンパラメトリックな MHDP によって推定された 22 を用いた．しかし，この上位カテゴリ数によっても形成される上位カテゴリは変化してしまう．そこで，KL ダイバージェンスを用い正解の同時確率と比較することで，上位カテゴリ数の妥当性について評価する．mMLDA により，上位カテゴリ数を変化させて概念形成を行い同時確率 $P(z^O, z^M)$ を計算し，正解となる同時確率 $\hat{P}(z^O, z^M)$ との KL 距離を計算した．その結果が図 3.12 であり，横軸

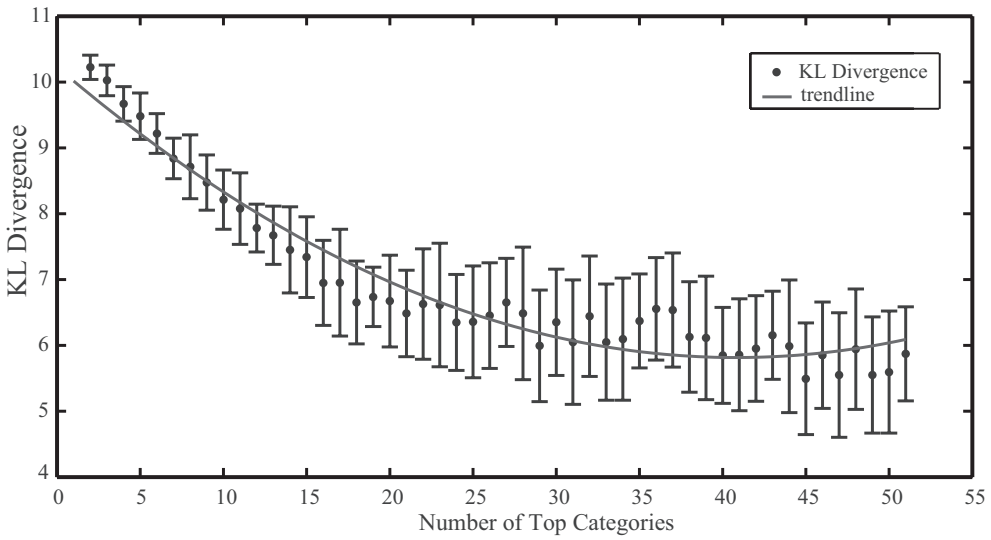

図 3.12: 上位カテゴリ数に対する同時確率分布の正解との KL ダイバージェンス

がカテゴリ数，縦軸が正解との KL ダイバージェンスである．カテゴリ数が少ない場合，少ないパラメータで物体と動きの関係を表現するため，正しく学習できず，正解との KL ダイバージェンスが大きくなっている．一方，カテゴリ数が多くなると，多くのパラメータで表現できるため，正しくその関係を捉えることができ，正解との KL ダイバージェンスが小さくなる．さらに，上位カテゴリ数が大きくなると，KL ダイバージェンスはほとんど変化しなくなるが，細かく分類しすぎてしまうために，正しい概念が形成できない恐れがある．そのため，図 3.12 より，上位カテゴリ数は 20〜30 の範囲が妥当であると考えられ，今回 MHDP で推定された上位カテゴリ数 22 は適切であると言える．

3.4.5 未観測情報の予測実験

次に，未観測情報の予測性能を評価するため，可観測の物体（動き）の情報から，未観測である動き（物体）概念の予測を行った．実験では，図 3.6 中の矩形で表示された物体を認識用データとして用い，残りの物体を学習用のデータとし，観測された物体のマルチモーダル情報（$\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h$）から動きカテゴリ z^M の予測を

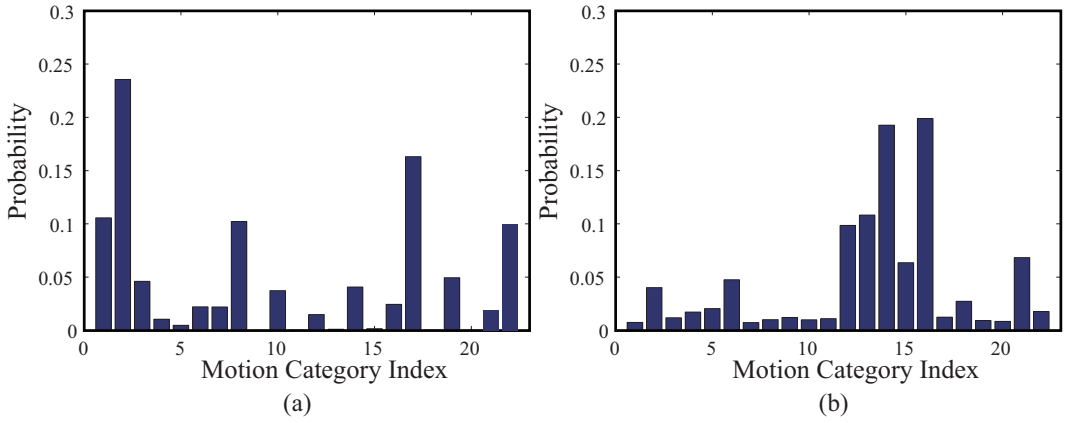

図 3.13: 「ぬいぐるみ（2)」から予測された動きの予測確率：(a）mMLDA，(b）近似モデル

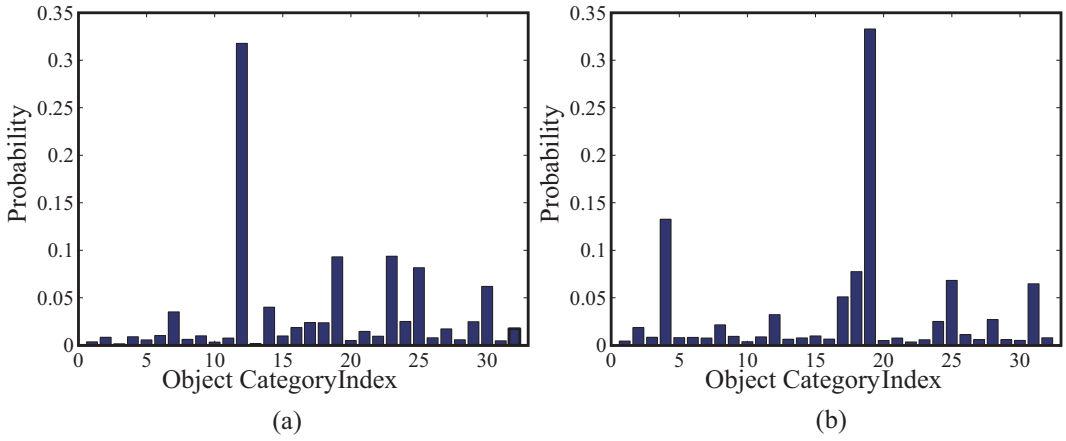

図 3.14: 「片手で口に運ぶ（3)」から予測された物体の予測確率：(a）mMLDA，(b）近似モデル

行った．同様に，観測された動きの情報 $\boldsymbol{w}^p$ から物体カテゴリ z^O の予測も行った．

その結果，mMLDA を用いた動きカテゴリ z^M の予測精度は 83.33%となり，近似モデルを用いた予測精度は 72.22%となった．同様に，観測された動きの情報から物体カテゴリ z^O を予測した結果，その予測精度は，mMLDA と近似モデルでそれぞれ，73.33%と 70.00%となった．mMLDA では，近似モデルに比べ，動きと物体の関係性が正しく捉えられているため高い精度となったと考えられる．

図 3.13 が,「ぬいぐるみ（2）」から予測された未観測である動きカテゴリが発生する確率 $P(z^M|\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h)$ である．mMLDA の結果では，正しく「上に投げる（2）」や「抱く（8）」といった動きを予測することができる．一方，近似モデルを用いた予測の結果では,「取り出す（13）」と「中身をかける（15）」といった動きが高い確率で予測されている．これは，近似モデルの分類結果では，物体の「ぬいぐるみ（2）」と「車（玩具）（28）」が同じカテゴリに分類されてしまったため，同じ上位カテゴリを持つ「フローリングワイパー（3）」に関係する「取り出す（13）」が予測されたと考えられる．このように，近似モデルでは，物体と動きが独立しており相互に影響を及ぼさないため，誤分類を修正することができず，予測精度が低下している．

また,「片手で口に運ぶ（3）」から予測された物体の発生確率 $P(z^O|\boldsymbol{w}^p)$ を図 3.14 に示した．mMLDA の結果では，正しく「金属のカップ（12）」を予測することができている．一方，近似モデルの結果では，誤った「スナック（19）」が最も高い確率で予測されている．これは,「片手で口に運ぶ（3）」と「食べる（9）」の動きが似通った情報を持ち，正しく認識できなかったことが原因であると考えられる．

以上のように，近似モデルに比べ mMLDA の予測性能が高いことが分かる．

3.5 まとめ

本章において，物体と動き概念とそれらの関係を表現する動作概念の獲得可能なモデルを提案した．提案モデル mMLDA は，確率モデルに基づくマルチモーダル LDA の多層化であり，下位概念とそれらを組み合わせた上位概念を形成する．実験結果より，提案した mMLDA が簡易的な近似モデルに比べ高い予測性能を持つことが明らかとなった．これは，上位・下位概念が相互に影響し合うことが，多

層概念形成において重要であることを物語っている．

提案した mMLDA は，人の動作に含まれている動きと物体の共起性を手がかりとして学習を行った．提案モデルは，前章で述べた確率的な知識の表現となっており，人の動作を観測することでロボット自身が知識を獲得することを可能にする．しかし，実際ロボットが獲得すべき知識は物体や動きだけではなく，例えば場所や人物など様々であり，mMLDA をより多様な概念に拡張する必要がある．さらに，本章では動作に関する概念は獲得できたが，それらの概念に対応する言語を獲得することはできていない．つまり，多様な概念の意味をどのように言葉として表現するのかを考える必要がある．この問題を解決するために，概念と言語（単語）を結び付ける手法や文法を獲得する枠組みが必要である．これにより，概念を基盤として，事物を文章として表現する仕組みについても考えることができる．これらの問題について，次章で議論する．

第4章　多様な概念を用いた言語獲得

4.1　はじめに

本章では，前章で提案した階層的カテゴリ分類のモデルである mMLDA を拡張し，より多様な概念を獲得することを考える．本章で提案する mMLDA は，前章と同様に先行研究で提案された MLDA [18] の階層構造となっており，下層の MLDA では下位概念である，物体，動き，場所，人物の概念がそれぞれ形成され，上層の MLDA ではこれらの概念を統合する上位概念が形成される．このモデルを用いることで例えば，下位概念としてジュースという物体概念や物を口に運ぶという動き概念，ダイニングという場所概念などが形成される．上位層ではこれらの関係性が学習され,「飲む」という行動概念が形成される．これにより，ジュースを見ることでそれを口に運ぶ「飲む」という行動や，その「飲む」という行動が「ダイニング」という場所で行なわれやすいといった未観測情報の予測を行うことが可能となる．また，形成された多様な概念を利用し，同時に語意や文法を獲得することで，観測したシーンを文章で表現する手法を検討する．先行研究においても，入力されたマルチモーダル情報に対応する単語や，単語が指す概念の推論の可能性が示されている [58]．しかし本章で扱う問題は，階層的な概念における語意の獲得であり，どの階層のどの概念にどの単語が結び付くかという問題を解く必要がある．この問題は，先行研究の手法では解くことができない．本章では，単語と概念間の相互情報量を用いることで，どの単語が本来どの概念に結び付いているのかを自動的に推定する手法を提案する．これにより単語と概念の結び付きを学習することが可能であり，各単語に対応する，物体，場所，人，動作といった概念クラスの推定が可能である．従って，教示発話における概念クラスの生起順を学習することで，概念クラスの遷移確率という形で表現される確率文法を学習す

ることが可能となる．

ここではさらに，このように獲得した文法と概念に結び付けられた語彙を用いて，観測されたシーンから文章を生成することを考える．このための一つ目の問題は，観測情報を表現するのに相応しい単語の推定であり，これは mMLDA を用いることで実現できる．この際に，未観測の情報を確率的に予測することが可能であり，これにより観測情報の曖昧性を過去の経験に基づいて解消できる可能性がある．二つ目の問題は，推定した単語の選択と語順の決定である．このために，学習した概念クラスの遷移確率を利用する．提案手法では，概念クラスの順番を確率文法からサンプリングすることで生成する．そして，各概念クラスと観測情報から表現に相応しい単語の選択を行う．また，文法として単語のバイグラムを用いることを考える．つまり，単語レベルでの遷移確率を考えることで，概念クラスの系列と各概念クラスに対応する単語候補のラティスから最も確率の高い単語系列を求める問題となり，これは Viterbi アルゴリズムを用いることで解くことができる．最終的に確率文法からのサンプリングを複数回行い，最も高い尤度の単語列を出力とする．提案したこれらのモデルを，実験によって評価する．

4.2 多様な概念の形成

本章では，MLDA を用いて形成された物体，動き，場所，人物の概念を統合することで，より上位の概念を階層的に形成することを考える．図 4.1 に提案する mMLDA のグラフィカルモデルを示す．図 4.1 において，z は統合概念を表すカテゴリであり，z^O，z^M，z^P，z^U はそれぞれ下位概念に相当する，物体，動き，場所，人物カテゴリである．上位カテゴリ z は，下位カテゴリ間の関係を表現したモデルとなっている．また，w^v, w^a, w^h は，それぞれ物体から得られる視覚，聴覚，触覚情報であり，w^p，w^c，w^s，w^y は物体を扱っている際の人の動き，座標，性別，年齢の情報である．さらに，w^w，w^{wO}，w^{wM}，w^{wP}，w^{wU} は，教示発話から得られる単語情報である．以下，下位及び統合概念について詳しく述べる．

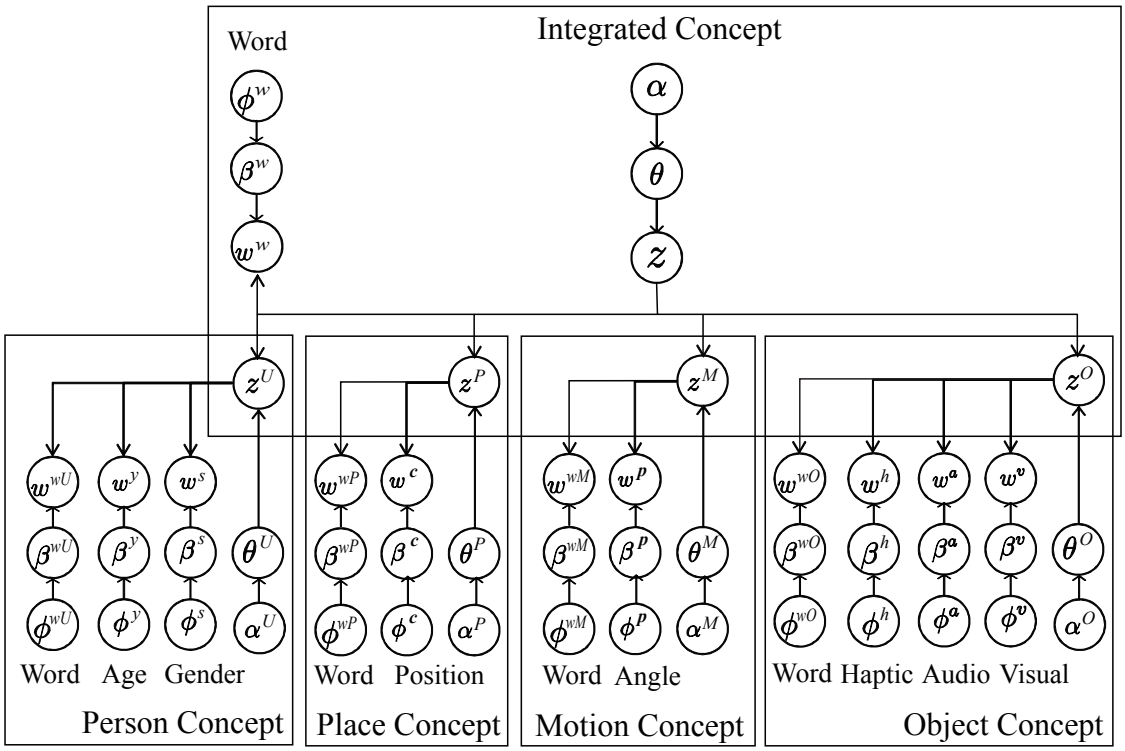

図 4.1: mMLDA のグラフィカルモデル

4.2.1 下位概念

物体概念は，ロボットが実際に取得したマルチモーダル情報をカテゴリ分類することにより形成する．つまり，視覚・聴覚・触覚・単語情報 w^v, w^a, w^h, w^{wO} がその類似性により分類され，物体カテゴリ z^O が形成される．ここで用いるロボットプラットフォームは，前章のアームロボット（図 3.4（a））と同じである．前述のように，知覚情報は Bag of Features（BoF）モデルを用いて表現する．視覚情報としては，取得した物体画像を 128 次元の DSIFT を用いて変換し，ベクトル量子化することで 500 次元のヒストグラムとする．聴覚情報は，MFCC を特徴量とし，ベクトル量子化することで 50 次元ヒストグラムとする．同様に，触覚情報は，取得した触覚センサのデータに対する曲線近似のパラメータをベクトル量子化し，15 次元ヒストグラムとする．また，全ての教示発話を単語分割し，Bag of Words（BoW）モデルを用いて表現したものを単語情報として扱う．

物体概念と同様に，図 4.1 の下側に示す MLDA と等価なモデルによって動き概念のモデル化を行う．動き情報は，人の動作中の 11 箇所の関節角度を，動作開始から動作終了まで KINECT を用いて取得することを前提とする．また動きの情報は，操作対象となる物体によって分節することができると仮定している．一つの動作から複数の 11 次元の特徴ベクトルが得られ，それをあらかじめ計算した 70 の代表ベクトルによりベクトル量子化することで 70 次元のヒストグラムとし，これを動き情報として用いる．

場所概念は，人の動作中の座標を動作開始から動作終了まで取得することで形成する．一つの動作から複数の 2 次元座標が得られるため，これらをベクトル量子化し，6 次元のヒストグラムとすることで場所情報とする．代表ベクトルは，学習データを K 平均法によりクラスタリングすることで決定する．

人物概念の形成では，動作中の人の顔画像から，性別及び年齢の推定を行い，これらの値を人物情報として扱う．他の概念と同様，性別・年齢の推定結果を基にデータの量子化を行い，2 次元の性別ヒストグラムと 10 次元の年齢ヒストグラムを人物情報として用いる．

4.2.2 統合概念

提案モデルにおいて，物体，動き，場所，人物概念は，それぞれ MLDA で表現されており，さらにその上位でそれらの概念の関係を統合概念としての MLDA で表現する二層構造となっている．物体，動き，場所，人物概念を独立した MLDA として学習し，その学習結果を統合概念として学習することも可能であるが，前章で明らかとなったように，そのような方法では各概念が相互に影響を与えることができない．そこで，各概念を表す隠れ変数 z，$z^C \in \{z^O, z^M, z^P, z^U\}$ を同時に学習する手法を以下に提案する．

学習にはギブスサンプリングを用いる．つまり，各概念を表すカテゴリ z，z^C を，観測データ $w^m \in \{w^v, w^a, w^h, w^{wO}, w^p, w^{wM}, w^c, w^{wP}, w^s, w^y, w^{wU}, w^w\}$ に基づいたサンプリングによって推定する．ただし，w^c はハイパーパラメータ ϕ^c によって決まるディリクレ事前分布に従う β^c をパラメータとする多項分布によって生成される．またカテゴリ z，z^C は，それぞれハイパーパラメータ α，α^C によって決ま

るディリクレ事前分布に従うパラメータ θ，θ^C をパラメータとする多項分布によって生成されるモデルである．サンプリングには，θ，θ^C，β^m を周辺化した以下の事後分布を用いる．

$$P(z_{jmi}, z^C_{jmi}|\boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{W}^m) \propto P(z_{jmi}|\boldsymbol{Z}_{-jmi})P(z^C_{jmi}|z_{jmi}, \boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi})P(w^m_{ji}|z^C_{jmi}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{W}^m_{-ji}) \quad (4.1)$$

なお，右辺のそれぞれの確率分布は次のようになる．

$$P(z_{jmi} = k|\boldsymbol{Z}_{-jmi}) = \frac{\alpha + N^{-jmi}_{j,z=k}}{K\alpha + N_j{}^{-jmi}}, \quad (4.2)$$

$$P(z^C_{jmi} = l|z_{jmi} = k, \boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}) = \frac{\alpha^C + N^{-jmi}_{z=k,z^C=l}}{K^C\alpha^C + N^{-jmi}_{z=k}}, \quad (4.3)$$

$$P(w^m_{ji} = x|z^C_{jmi} = k, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{W}^m_{-ji}) = \frac{\phi^m + N^{-jmi}_{z^C=k,w^m=x,m}}{W^m\phi^m + N^{-jmi}_{z^C=k,m}}, \quad (4.4)$$

ただし，$\boldsymbol{Z}$，$\boldsymbol{Z}^C$ は，それぞれ全物体の全情報に割り当てられた上位カテゴリと下位概念のカテゴリの集合を表し，$\boldsymbol{W}^m$ はモダリティ m の全物体の情報の集合である．N_{jz} は物体 j の全モダリティに上位カテゴリ z が割り当てられた回数であり，$N_{z^C w^m}$ はモダリティ m の特徴量 w^m に下位カテゴリ z^C が割り当てられた回数である．また，N_{z,z^C} は上位カテゴリ z と下位カテゴリ z^C の共起した回数を表しており，K，K^C，W^m はそれぞれ上位カテゴリのカテゴリ数，概念 C のカテゴリ数，モダリティ m の情報の次元数である．負の添字はその情報を除外することを表し，$-jmi$ は j 番目の物体のモダリティ m の i 番目の情報を除外することを表している．

モデルの学習は，隠れ変数である z，z^C を，収束するまで事後分布からサンプリングすることによって実現できる．しかし，隠れ変数が複数あり，複雑なモデルであるため，全てのパラメータを同時に求めると局所解に陥りやすいといった問題がある．そこで前章と同様に，図 4.1 の右側に示す下位カテゴリ z^C を個々の独立した MLDA として学習し，下位概念のパラメータ β^m を先に決定する．この

とき，各カテゴリ $z^C \in \{z^O, z^M, z^P, z^U\}$ は，次式を用いてサンプリングする．

$$\begin{aligned} z^C_{jmi} &\sim P(z^C_{jmi}|w^m_{ji}, \boldsymbol{W}^m_{-ij}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi}) \\ &\propto \sum_z P(z|\boldsymbol{Z}_{-jmi}) P(z^C_{jmi}|\boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}, z) \\ &\quad \times P(w^m_{ji}|\boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, z^C_{jmi}) \end{aligned} \tag{4.5}$$

このサンプリングを収束するまで繰り返すことで，式（4.4）を決定する．次に，式（4.4）を固定し，上位カテゴリ z，下位カテゴリ z^C をサンプリングする．

$$\begin{aligned} z_{jmi} &\sim P(z_{jmi}|w^m_{ji}, \boldsymbol{W}^m_{-ij}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi}) \\ &\propto \sum_{z^C} P(z_{jmi}|\boldsymbol{Z}_{-jmi}) P(z^C|\boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}, z_{jmi}) \\ &\quad \times P(w^m_{ji}|\boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, z^C) \end{aligned} \tag{4.6}$$

このとき，下位カテゴリ z^C が上位概念の影響を受けて更新されることに注意が必要である．Algorithm 3 と Algorithm 4 がそれぞれ，下位概念のパラメータの決定と，モデル全体の学習アルゴリズムである．以上のようなサンプリングを繰り返すことで，N_* がある値へと収束する．K を上位カテゴリのカテゴリ数とするとき，最終的なパラメータの推定値 $\hat{\beta}^m_{w^m z^C}$，$\hat{\theta^C}_{zz^C}$，$\hat{\theta}_{jz}$ は以下のようになる．

$$\hat{\beta}^m_{w^m z^C} = \frac{N_{z^C w^m m} + \phi^m}{N_{z^C m} + W^m \phi^m}, \quad \hat{\theta^C}_{zz^C} = \frac{N_{zz^C m} + \alpha^C}{N_{zm} + K\alpha^C}, \quad \hat{\theta}_{jz} = \frac{N_{jz} + \alpha}{N_j + K\alpha}, \tag{4.7}$$

ただし，W^m はモダリティ m の次元数を表し，$N_{z^C w^m m}$ はモダリティ m の w^m に下位カテゴリ z^C が割り当てられた回数を表す．

Algorithm 3 Multilayered MLDA (bottom layer)

for all i, j, C, m **do**
  $u \leftarrow$ draw from Uniform [0,1]
  for $k \leftarrow 1$ to K^C **do**
    $P[k] \leftarrow P[k-1] + P(z^C_{jmi} = k | w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi})$
  end for
  for $k \leftarrow 1$ to K^C **do**
    if $u < P[k]/P[K^C]$ **then**
      $z^C_{jmi} = k$, break
    end if
  end for
end for

Algorithm 4 Multilayered MLDA (whole layer)

for all i, j, C, m **do**
  for $k \leftarrow 1$ to K **do**
    $P[k] \leftarrow P[k-1] + P(z_{jmi} = k | w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi})$
  end for
  $u \leftarrow$ draw from Uniform [0,1]
  for $k \leftarrow 1$ to K **do**
    if $u < P[k]/P[K]$ **then**
      $z_{jmi} = k$, break
    end if
  end for
  for $k \leftarrow 1$ to K^C **do**
    $P[k] \leftarrow P[k-1] + P(z^C_{jmi} = k | w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi})$
  end for
  $u \leftarrow$ draw from Uniform [0,1]
  for $k \leftarrow 1$ to K^C **do**
    if $u < P[k]/P[K^C]$ **then**
      $z^C_{jmi} = k$, break
    end if
  end for
end for

4.3 未観測情報の予測

学習したモデルを用いることで，物体や動きの認識だけでなく，概念間の予測も可能となる．例えば，場所概念 $\hat{z}^P$ と上位カテゴリ z は観測されたモダリティ v，

a，h（物体），p（動き）$\boldsymbol{w}_{\mathrm{obs}}^{v,a,h,p}$ を以下の式より推定することができる．

$$P(z, z^O, z^M|\boldsymbol{w}_{\mathrm{obs}}^{v,a,h,p}) =$$
$$P(z)P(z^O|z)P(z^M|z)P(\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h|z^O)P(\boldsymbol{w}^p|z^M), \tag{4.8}$$
$$\hat{z}^P = \underset{z^P}{\mathrm{argmax}} \sum_{z,z^O,z^M} P(z^P|z)P(z, z^O, z^M|\boldsymbol{w}_{\mathrm{obs}}^{v,a,h,p}), \tag{4.9}$$
$$\hat{z} = \underset{z}{\mathrm{argmax}} \sum_{z^O,z^M} P(z, z^O, z^M|\boldsymbol{w}_{\mathrm{obs}}^{v,a,h,p}) \tag{4.10}$$

さらに，形成された概念を利用することで未観測情報の予測を行うことも可能である．例えば，学習したモデルを用いて物体概念の単語は観測されたモダリティ v，a，h から次式より予測することができる．

$$P(w^{wO}|\boldsymbol{w}_{\mathrm{obs}}^{v,a,h}) = \int \sum_{z^O} p(w^{wO}|z^O)p(z^O|\theta^O)p(\theta^O|\boldsymbol{w}_{\mathrm{obs}}^{v,a,h})d\theta^O \tag{4.11}$$

4.4 近似モデル

前述のように，全ての下位概念は，統合概念を無視すれば，独立した MLDA と等価なモデルと考える事ができる．さらに，単語情報 w^w 以外の w^m を無視することで，統合概念は z^O，z^M，z^P，z^U と w^w を生成する MLDA と等価なモデルと見なすことができる．すなわち各概念を独立した MLDA として学習し，フィードフォワード的に接続することで，簡易的に多様な概念を統合することができる．図 4.2 が mMLDA を分解し，独立した 5 つの MLDA として考えた場合のグラフィカルモデルである．このモデルでは，下位カテゴリ $z^C \in \{z^O, z^M, z^P, z^U\}$ を独立した MLDA で学習した後，上位カテゴリ z をもう一つの独立な MLDA で学習すればよい．

近似モデルの学習ではまず，下位層を独立した MLDA として学習した後に，z^C を多項分布 $P\left(z^C \mid \boldsymbol{w}^{m_C}\right)$ からサンプリングする．ただし，$\boldsymbol{w}^{m_C}$ は概念 C のモダリティを表している．近似モデルの上位層に相当する MLDA では，生成された z^C をそれぞれ図 3.2 に示した w^* として考えることで学習することができる．従って，下位概念の関係性は，モデルにおける隠れ変数 z によって学習され，この z が下

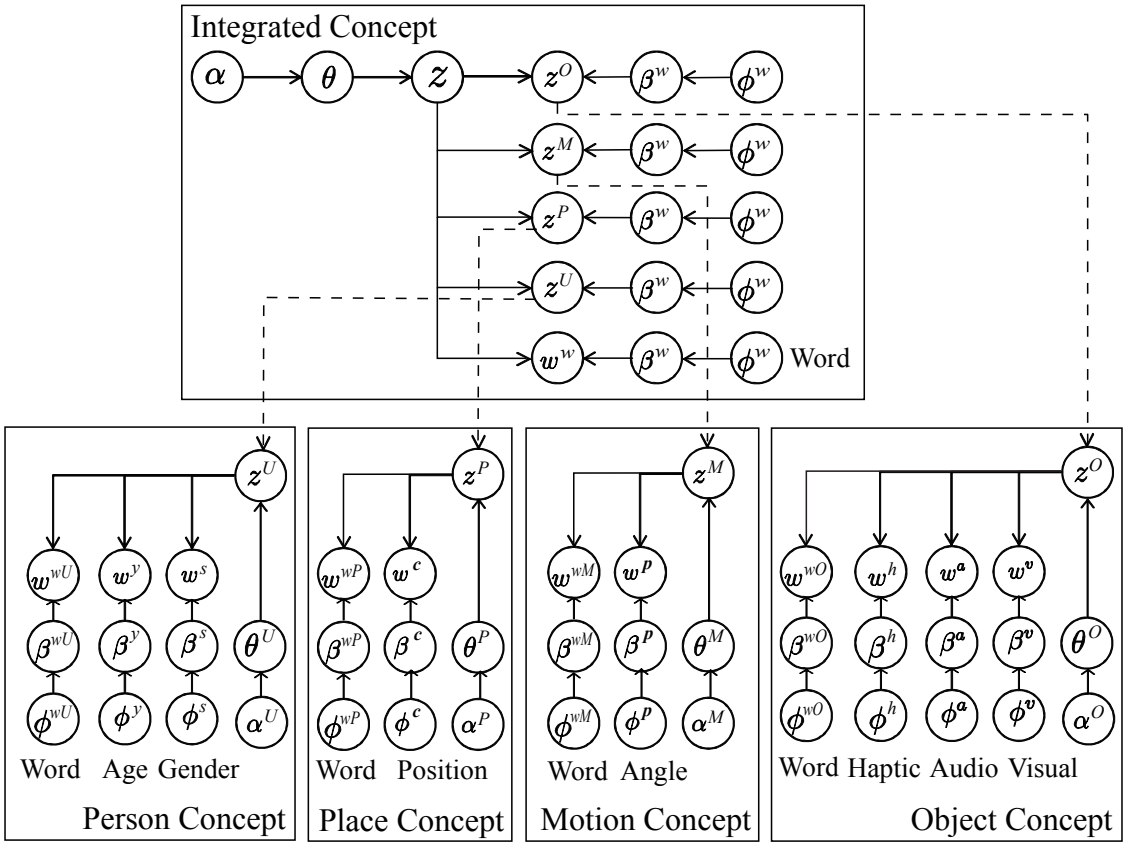

図 4.2: 近似多層マルチモーダル LDA のグラフィカルモデル

位概念の統合的な概念を表現するカテゴリとなる．ただし，z は固定された z^C の関係性を表現するだけであり，逆に z^C に影響を与えることはない．

近似モデルにおいても，学習したモデルを用いて未観測情報を予測することが可能である．例えば，人物情報 $\boldsymbol{w}^m$ から，確率の高い場所カテゴリを次のように予測することができる．

1. 人物情報から人物カテゴリをサンプリングする

$$\hat{z}^U \sim P\left(z^U \mid \boldsymbol{w}^y, \boldsymbol{w}^s\right) \tag{4.12}$$

2. 次式によりサンプリングされた人物カテゴリ $\hat{\boldsymbol{z}}^U$ から，場所カテゴリが発生

する確率を計算する

$$P(\hat{z}^P|\hat{\boldsymbol{z}}^U) = \int \sum_z P(\hat{z}^P|z)P(z|\theta)P(\theta|\hat{\boldsymbol{z}}^U)d\theta \tag{4.13}$$

さらに，$\hat{z}^P$ に $\hat{z}^O$，$\hat{z}^M$ を代入することで，他の概念のカテゴリも予測することができる．

前章で述べたように近似モデルの利点は，モデルがシンプルで実装し易いことである．欠点は，上位概念から下位概念に影響を与えずに独立に学習を行うため，下位層に生じた分類の誤りはそのまま統合概念の学習に影響を及ぼしてしまい，後の実験結果から分かるように，相互に学習を行う mMLDA に比べてモデル全体の精度を下げることにつながることである．本章では，近似モデルをベースラインモデルとして，提案する mMLDA を比較する．

4.5 言語学習

4.5.1 相互情報量を用いた単語の予測

本章では，図 4.1 に示したように，各概念に教示発話から得られる全ての単語情報を与えて学習を行う．各概念を表現する適切な単語が存在すると考え，単語とカテゴリの結び付きの強さの尺度として，単語とカテゴリ間の相互情報量を用いる．単語 w^w と概念クラス $i \in$ { 物体概念，動き概念，場所概念，人物概念，統合概念 } のカテゴリ k との間の相互情報量は以下の式となる．

$$I(w^w, k|i) = \sum_{K,W} P(W, K|i) \log \frac{P(W, K|i)}{P(W|i)P(K|i)} \tag{4.14}$$

ただし，$K \in (k, \bar{k})$，$W \in (w^w, w^{\bar{w}})$ とし，$\bar{k}$ は k 以外のカテゴリを表す．また，$w^{\bar{w}}$ は w^w 以外の単語を表している．相互情報量とは，二つの確率変数の共有する情報量であり，相互依存の尺度である．したがって単語とカテゴリ間の相互情報量が大きい場合，その単語はそのカテゴリを表現していると言える．

本章では，単語によって，複数の概念を表す可能性があると考え，式（4.14）を

用いて求めた相互情報量を単語の各概念クラスに対する重みとして考える．その重みを $F(i, w^w)$ とし，次式で単語予測スコアを計算する．

$$F(i, w^w) = \max_k I(w^w, k|i) \tag{4.15}$$

$$\hat{P}(w^w|\boldsymbol{w}^m_{\rm obs}, i) = F(i, w^w) P(w^w|\boldsymbol{w}^m_{\rm obs}, i) \tag{4.16}$$

このように，単語の各概念クラスに対する重みを求め，概念クラス i の単語予測 $P(w^w|\boldsymbol{w}^m_{\rm obs}, i)$ の際に重みを付けることで，各概念から生成される単語の予測精度を向上させることが可能となる．

4.5.2 文法の学習

mMLDA を用いることで，観測情報を表現するのに適切な単語を予測することができる．文章を生成するためには，さらに文法を考える必要がある．本章では，mMLDA における概念クラス（式（4.16）における i）の発火順を文法と考える．これは，各単語に対する概念選択により，単語は特定の概念クラスと結び付けることができるためであり，教示発話を単語分割しその各単語の概念クラスを推定することで実現できる．ただし，ここでは助詞や機能語を考えないこととする．

例えば,「母はキッチンで野菜を切る」という発話から概念クラスを,「母－人物 ; キッチン－場所 ; 野菜－物体 ; 切る－動き」と推定することができ，結果的に「(人物)（場所）（物体）（動き)」となる文法が得られる．ここでは文法を，マルコフモデルで表現することとし，教示発話から次のように学習する．

$$P(C_t|C_{t-1}) = \frac{N_{C_{t-1}C_t}}{N_S} \tag{4.17}$$

ただし，C_t は文章中の t 番目の単語に該当する概念クラスである．また，$N_{C_{t-1}C_t}$ と N_S はそれぞれ C_{t-1} から C_t に遷移した回数と概念クラス間遷移の総数である．

4.6 観測情報からの文生成

4.6.1 概念遷移に基づく文生成

ロボットが観測情報を文章で表現するための最も単純な方法として，予測した単語を上記の文法をもとに並べればよい．具体的には Algorithm 5 に示すように，文頭 BOS から順に $P(C_t|C_{t-1})$ に従って t 番目の概念クラスである C_t をサンプリングする．そして，$w_t = \mathrm{argmax}_{w^w} \hat{P}(w^w|\boldsymbol{w}^m_{\mathrm{obs}}, C_t)$ に従って t 番目の単語を計算する．この手順を，C_t が文末 EOS になるまで繰り返す．この手法（以下「Method 1」と呼ぶ）では，サンプリングされた概念系列 $\boldsymbol{C} = \{C_0 = BOS, C_1, \cdots, C_T = EOS\}$ に対して，単語 w_t を埋めていくことで文生成するため，$\boldsymbol{C}$ に同じ概念がサンプリングされてしまうと，同じ単語で文が生成され，不自然な文となる．

4.6.2 言語モデルを用いた文生成

本章ではより自然な文を生成するために，言語モデルを導入し，文生成を確率的に定式化する（以下「Method 2」と呼ぶ）．まず，Method 1 と同様，式（4.17）に従い，BOS から EOS までの概念系列を N 個サンプリングし，n 番目の BOS と EOS を除いたサンプルを $\boldsymbol{C}^n = \{C^n_1, \cdots, C^n_t, \cdots, C^n_{T_n-1}\}$ とする．次に，概念 C^n_t から，概念と対応した単語を生成する．ここでは，観測情報 $\boldsymbol{w}^m_{\mathrm{obs}}$ が与えらたとき，概念 C^n_t と対応した単語の発生確率の高い上位 Q 個の単語 $\boldsymbol{w}^n_t = \{w^n_{t1}, w^n_{t2}, \cdots, w^n_{tQ}\}$ を用い，全ての単語の集合を $\boldsymbol{W}^n = \{\boldsymbol{w}^n_1, \boldsymbol{w}^n_2, \cdots, \boldsymbol{w}^n_{T_n-1}\}$ とする．すなわち，これらの概念系列・単語から Q^{T_n-2} パターンの文を生成することができ，文 S^n が生成される確率を次のように定義する．

$$P(S^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{\mathrm{obs}}) \propto \prod_t P(C^n_t|C^n_{t-1}) P(w^n_t|\boldsymbol{w}^m_{\mathrm{obs}}, C^n_t) P(w^n_t|w^n_{t-1}) \qquad (4.18)$$

次に，観測情報に対してサンプリングされた N 個の概念系列と単語から生成した文の中から，生成される確率が高い文を選択する．まず，各概念系列に対して，各概念系列から，式（4.18）を最大とする文 $\hat{S}^n$ を決定する．ただし，S^n のパターンは非常に多く，単純には決定することができないため，Viterbi アルゴリズムを用

Algorithm 5 Concept Transition based Sentence Generation

$t \leftarrow 1, C_0 = BOS$
for t **do**
  $C_t \leftarrow$ draw from $P(C_t|C_{t-1})$
  if $C_t = EOS$ **then**
    break
  end if
  $w_t = \mathrm{argmax}_{w^w} \hat{P}(w^w|\boldsymbol{w}^m_{\mathrm{obs}}, C_t)$
  $t \leftarrow t + 1$
end for

Algorithm 6 Language Model and Concept Transition based Sentence Generation

for $n \leftarrow 1$ to N **do**
  $t \leftarrow 1, C^n_0 = BOS$
  for t **do**
    $C^n_t \leftarrow$ draw from $P(C^n_t|C^n_{t-1})$
    if $C^n_t = EOS$ **then**
      break
    end if
    $\boldsymbol{w}^n_t \sim \hat{P}(w^w|\boldsymbol{w}^m_{\mathrm{obs}}, C^n_t)$
    $t \leftarrow t + 1$
  end for
  $\hat{S}^n = \mathrm{argmax}_{S^n} P(S^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{\mathrm{obs}})$
end for

いて式（4.18）の確率が最大となる文を探索する．ここで，各概念系列のサンプルに対して確率が最大となる文の集合を $\hat{\boldsymbol{S}} = \{\hat{S}^0, \cdots, \hat{S}^n, \cdots, \hat{S}^N\}$ とする．以上の手順を Algorithm 6 にまとめる．

次に，$\hat{\boldsymbol{S}}$ から生成される確率が最大となる文を選択することで，最終的な生成文とする．しかし，実際には文が長いほど，その確率は小さくなってしまう．そこで以下のような，調整係数 $\ell(\hat{S}^n)$ を導入する．

$$\ell(\hat{S}^n) = \frac{(L^{\max} - L_{\hat{S}^n})}{\sum_n^N L_{\hat{S}^n}} \sum_n^N P(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{\mathrm{obs}}) \tag{4.19}$$

ただし，$L_{\hat{S}^n}$ は $\hat{S}^n$ の文の長さ，$L^{\max}$ は $\hat{\boldsymbol{S}}$ の中の文長の最大値である．式（4.19）を用いて，文の確率を次のように再定義する．

$$\bar{P}(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{\text{obs}}) = P(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{\text{obs}}) + \ell(\hat{S}^n) \tag{4.20}$$

従って最終的な文 S は，以下のように求める．

$$S = \underset{\hat{S}^n \in \hat{\boldsymbol{S}}}{\text{argmax}}\, \bar{P}(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}^m_{\text{obs}}) \tag{4.21}$$

4.7 実験

提案モデルの有効性を検証するために，実験を行った．実験に用いたデータセットの一例を図 4.3 に示す．データセットは 132 個の物体を用い，正解として 32 個のカテゴリに分けた．これらの物体を使用する動作を行い，実験のためのデータを取得した．全ての物体を，概念形成の実験に使用した．予測実験においては，各カテゴリから一個の物体（図 4.3 の赤枠）をテストセットとして使用し，残りの物体を学習データとする．物体概念は観測されたマルチモーダル情報から形成された．動き情報として，KINECT を用いて人の各関節の角度を取得した．動き情報の一例と KINECT より取得された骨格情報を図 4.3 に示す．本実験においては，人物のデータとして，大人・子供の男女の画像をインターネットからダウンロードし，OKAO Vision [63, 64] が提供する画像センシング技術による年齢・性別推定を使用した．図 4.3 に使用した画像及びヒストグラム化した性別と年齢推定の結果を示す．さらに，場所の情報としては，図 4.3 に示したような家の間取りを仮定し，玄関，リビング，キッチン，ダイニング，浴室，庭の 6 つの場所を想定し，座標のデータを取得した．全てのデータの組み合わせを表 4.1 に示す．

本章では人が表 4.1 に示した各データに対して，5 つの教示発話を与えることとする．取得した全ての教示発話は形態素解析器を用いて単語分割を行い，他の知覚情報と同様に BoW モデルとしてヒストグラム化し生起回数の情報として取り扱う．表 4.2 に教示発話の一例を示す．

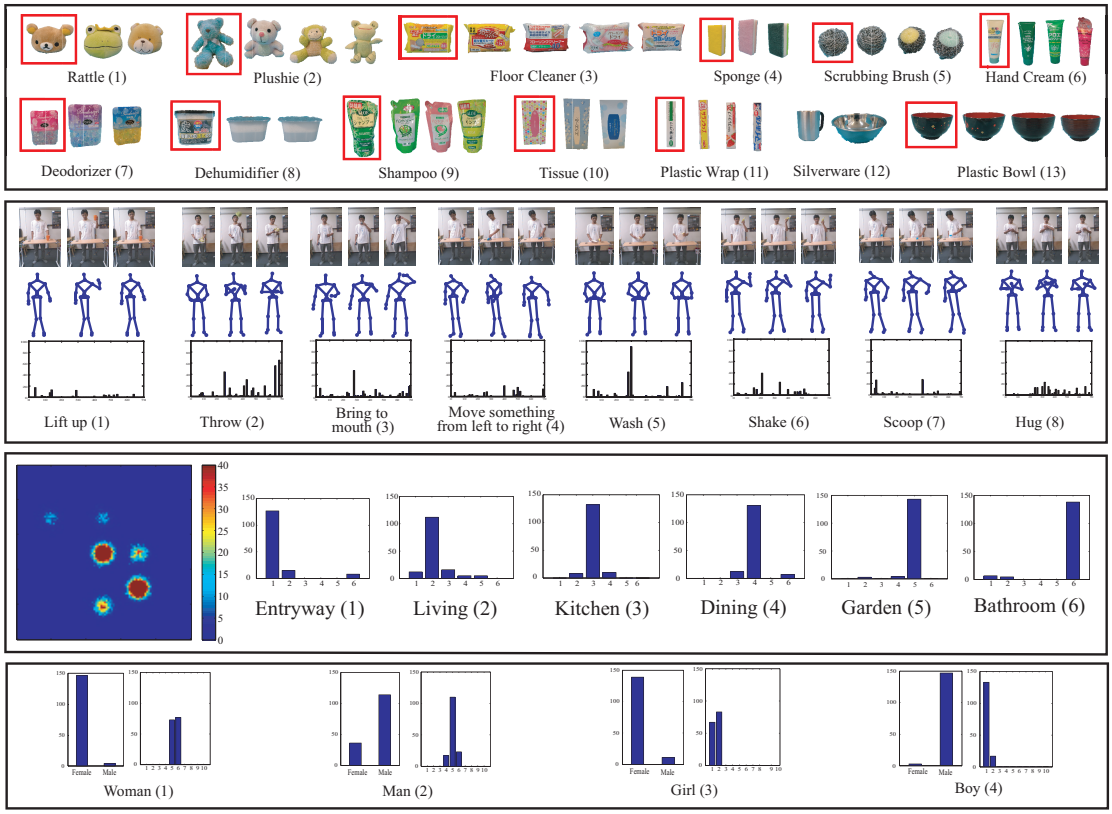

図 4.3: 実験に用いたデータセットの例．最上のボックスが物体の例を示し，赤枠が認識実験に用いた物体を表す．二番目のボックスが取得した動き情報：（上から下まで）物体に対して行った動きの例，（上）KINECT の画像，（中）実際の動き，（下）70 次元のヒストグラム．三番目のボックスが，場所全体における位置の集中分布（左），場所情報（右）を示し各場所に対して 6 次元のヒストグラム．最下のボックスが人物情報の例を示し，各概念に対して，2 次元の性別情報（左），10 次元の年齢情報（右）．2〜4 番目のボックスの括弧内の番号はカテゴリ番号を表す

4.7.1 カテゴリ数決定

カテゴリ数をどのように決定するかについては，前章と同様の手法を用いることができる．まず下位層については，ノンパラメトリックベイズ手法であるマルチモーダル階層ディリクレ過程（Multimodal Hierarchical Dirichlet Process: MHDP）[62] による決定手法をそのまま利用する．実際に MHDP によって下位層のカテゴリ数を推定したところ，図 4.4 に示したように物体，動き，場所，人物のカテゴリ数は

表 4.1: 動き，物体，場所，人物データの対応表（カッコ内の数字はカテゴリ ID）

動き	物体	場所	人物
持ち上げる（1）	茶碗（13）	ダイニング（4）	全員（1，2，3，4）
	飲み物（缶）（17）		
	カップヌードル（21）		
	プラスチックカップ（25）		
	スプレー缶（23）	庭（5）	男性（2，4）
上に投げる（2）	ぬいぐるみ（2）	リビング（2）	子供（3，4）
	マラカス（29）		
	ボール（31）		
口に運ぶ（3）	金属の食器（12）	ダイニング（4）	全員（1，2，3，4）
	飲み物（缶）（17）		
	ペットボトル（18）		
	プラスチックカップ（25）		
	茶碗（13）		
	野菜（玩具）（27）		
	カップヌードル（21）		
	スナック（19）	リビング（2）	
左右に動かす（4）	車（玩具）（28）		
	フリーリングワイパー（3）	ダイニング（4）	大人の女性（1）
皿を洗う（5）	スポンジ（4）	キッチン（3）	
	たわし（5）		
上下に振る（6）	ガラガラ（1）	リビング（2）	子供（3，4）
	マラカス（29）		
	ドレッシング（14）	ダイニング（4）	全員（1，2，3，4）
	ソース（16）		
	飲み物（缶）（17）		
	ペットボトル（18）		
	スプレー缶（23）	庭（5）	大人の男性（2）
すくう（7）	ショベル（26）		
抱く（8）	ぬいぐるみ（2）	リビング（2）	女の子（3）
積み重ねる（9）	積み木（32）		子供（3，4）
置く（10）	消臭剤（7）		大人の女性（1）
	除湿剤（8）		
	積み木（32）		子供（3，4）
	プラスチックカップ（25）	ダイニング（4）	
手に塗る（11）	ハンドクリーム（6）	リビング（2）	女性（1，3）
取り出す（12）	ティッシュ箱（10）		全員（1，2，3，4）
	クッキー（20）		
	フローリングワイパー（3）	ダイニング（4）	大人の女性（1）
ナイフで切る（13）	野菜（玩具）（27）	キッチン（3）	
中身をかける（14）	ドレッシング（14）	ダイニング（4）	全員（1，2，3，4）
	蜂蜜（15）		
	ソース（16）		
中身を注ぐ（15）	シャンプー（9）	浴室（6）	大人（1，2）
	じょうろ（24）	庭（5）	大人の男性（2）
	飲み物（缶）（17）	ダイニング（4）	全員（1，2，3，4）
	ペットボトル（18）		
包む（16）	ラップ（11）		大人の女性（1）
塗る（17）	スプレー缶（23）	庭（5）	大人の男性（2）
履く（18）	靴（30）	玄関（1）	全員（1，2，3，4）
袋を開ける（19）	スナック（19）	リビング（2）	

表 4.2: 教示発話の例

教示発話
女の子はリビングで腕を上下に動かしてガラガラを振って音を聞く
女の子はリビングでガラガラを上下に動かして、振って音を聞く
父は浴室でシャンプーをもって、中身を注いで詰め替える
母はダイニングでピンク色のフローリングワイパーを開けて中身を取り出す
女の子はリビングでぬいぐるみを上に投げて遊んでいる

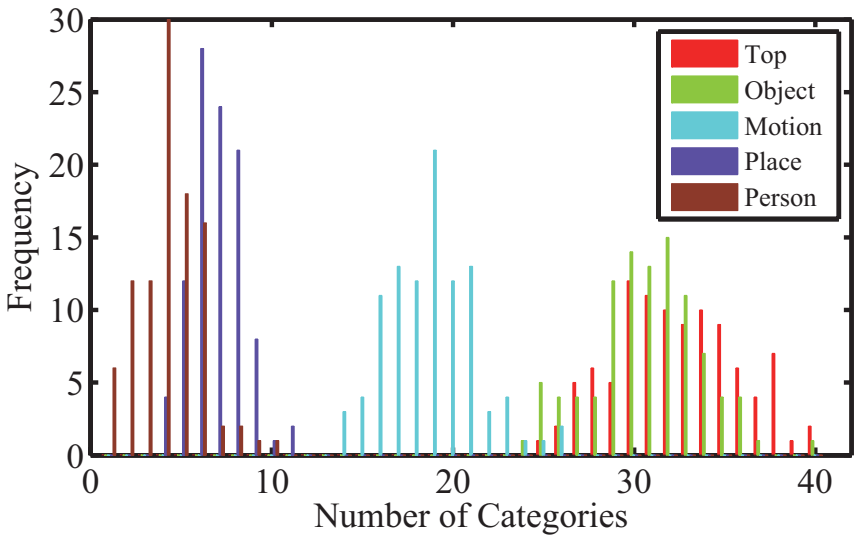

図 4.4: MHDP を用いた各概念のカテゴリ数の発生頻度

それぞれ 32，19，6，4 と推定された．この結果を用いて以降の実験を行うと共に，上位層のカテゴリ数を推定するためにも用いる．

上位層のカテゴリ数は，MHDP を直接適用して推定することができないため，近似モデルの上位 MLDA に MHDP を適用することでカテゴリ数を推定する．MHDP はサンプリングにより学習を行っているため，初期値によって推定されるカテゴリ数が変わってしまう．そこで，MHDP を用いた分類を 100 回行い，100 個のモデルを学習した．図 4.4 が 100 個のモデルのカテゴリ数のヒストグラムであり，横軸と縦軸はそれぞれ，推定した上位カテゴリ数とその頻度を示している．すなわち，このグラフはカテゴリ数の発生確率と考えることができ，カテゴリ数 30 が最も高い確率で発生していることが分かる．

以上の結果から，上位カテゴリ数を 30，物体，動き，場所，人物のカテゴリ数はそれぞれ 32，19，6，4 として，mMLDA と近似モデルによって概念形成を行い，各概念の評価を行った．

4.7.2 下位概念

ここでは，mMLDA と近似モデルによって形成された下位概念である物体，動き，場所及び人物概念をそれぞれ評価し，それらの結果について説明する．まず，物体概念の形成結果について述べる．物体概念の形成結果は図 4.5 であり，縦軸が物体のカテゴリ番号，横軸がモデルによって分類されたカテゴリを表している．図 4.5（a）が人手による分類であり，これを正解として各手法の分類結果を評価した．図 4.5（b）が mMLDA による分類結果であり，図 4.5（c）が近似モデルによる分類結果である．これらの分類結果から，図 4.5（a）を正解として，前章で定義した式（3.28）により分類精度を計算した．分類精度を計算した結果，mMLDA では 74.24%，近似モデルでは 65.15%となり，mMLDA の方がより正解に近い分類ができている．形成された物体概念の例として,「飲み物（缶）(17)」を取り上げて比較する．近似モデルでは，この物体を 3 つのカテゴリに分けてしまっているが，mMLDA では一つのカテゴリに分類された.「飲み物（缶）(17)」に属する物体は，異なるテクスチャ（知覚情報）を持っているため，この手がかりのみを用いて分類する近似モデルでは複数のカテゴリに分類されてしまった．また本実験において，物体概念のカテゴリ数は前章と同じ数に設定されているため，概念間の関係を考慮しない近似モデルでは，前章と同じ結果が得られることになる．一方 mMLDA では，物体の知覚情報のみならず，概念間の関係を考慮するため，正しく分類することが可能となった．本章の結果が前章の結果と同じであるのは,「飲み物（缶）(17)」に対して与えられた情報の組合せが前章と同じためである．まず「飲み物（缶）(17)」に対して，動きの組合せは前章と本章は同じである．また，表 4.1 より,「飲み物（缶）(17)」における場所と人物の組合せはどれも同じである．つまり,「飲み物（缶）(17)」という概念は，同じ手がかりで形成されたことになる．

次に，mMLDA と近似モデルによって分類された動き概念を評価した．図 4.6 が分類結果であり，縦軸が実際の動きのカテゴリ番号，横軸が分類されたカテゴリ

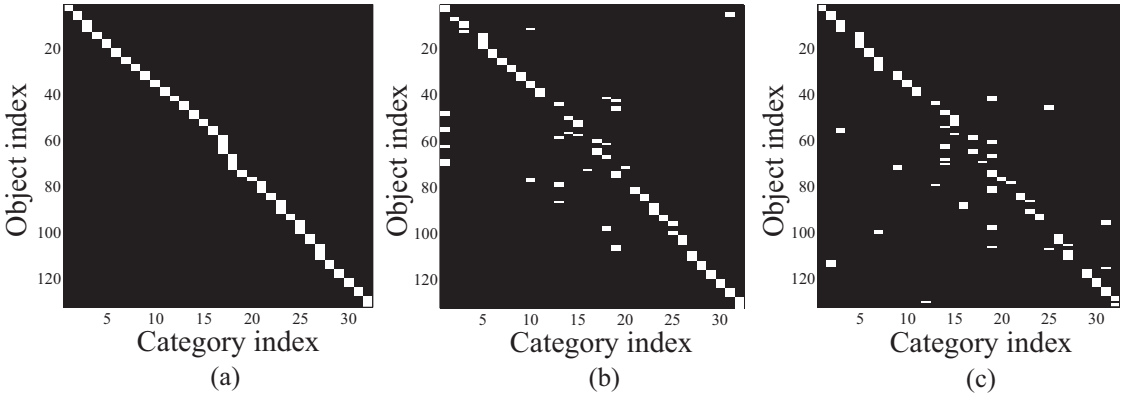

図 4.5: 物体の分類結果：(a) 正解, (b) mMLDA, (c) 近似モデル

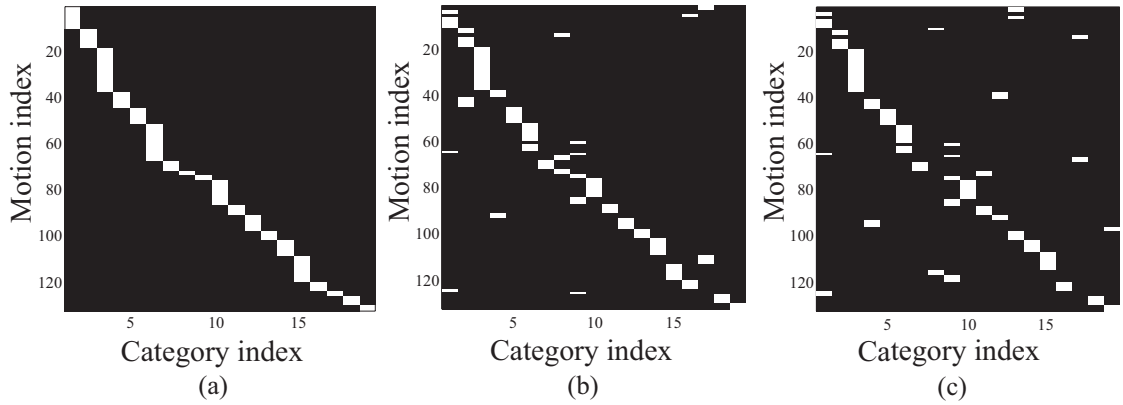

図 4.6: 動きの分類結果：(a) 正解, (b) mMLDA, (c) 近似モデル

番号である．図 4.6 (a) が人手による分類であり，物体と同様，この分類を正解として，各種法の分類を評価した．図 4.6 (b) が mMLDA による分類結果，図 4.6 (c) が近似モデルによる分類結果である．正解の分類（図 4.6 (a)）と比較すると，mMLDA（図 4.6 (b)）の分類精度は 81.06%となり，近似モデル（図 4.6 (c)）の分類精度は 75.00%となった．mMLDA と近似モデルによる動き概念の形成結果の差異は,「中身をかける (14)」の分類結果で見ることができる．mMLDA の分類結果では，この動きを一つに分類することができた．一方，近似モデルではこの動きを二つのカテゴリに分類してしまい，一部は「中身を注ぐ (15)」と同一のカテゴリとなった．これらの分類結果に対する要因として，物体概念の例と同じよう

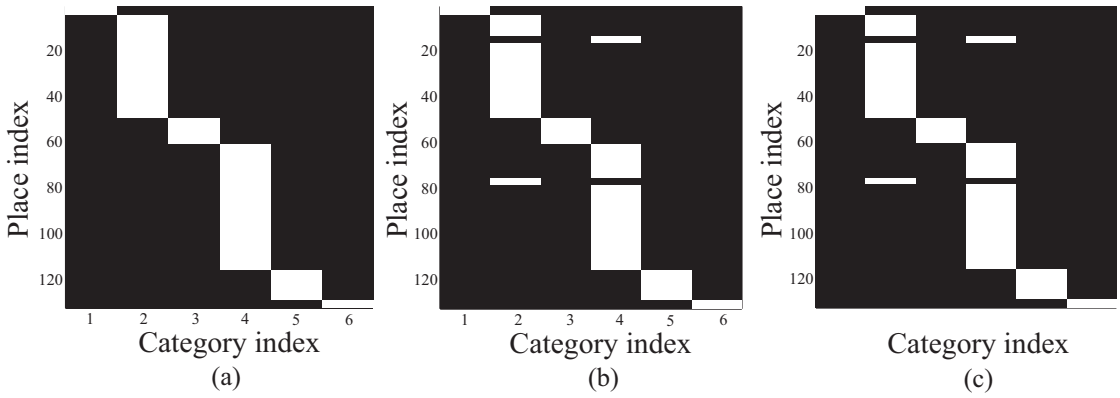

図 4.7: 場所の分類結果：(a) 正解, (b) mMLDA, (c) 近似モデル

に似通った知覚情報のみによる分類が困難なことが挙げられる．この場合，似た動き情報だけを手がかりとして分類する近似モデルでは動き概念を細かくしてしまった．これに対して，動き概念と関係する物体や場所なども手がかりとして概念を形成する mMLDA では，正しく一つのカテゴリに分類することができている．

物体と同様，図 4.7 が人手による場所概念の分類であり，この分類を正解として，mMLDA と近似モデルの分類を評価した．図 4.7（b）が mMLDA による分類であり，図 4.7（c）が近似モデルによる分類結果である．正解の分類（図 4.7（a））と比較すると，mMLDA（図 4.7（b））と近似モデル（図 4.7（c））の分類精度は共に 96.97%であった．本実験で用いた場所に関するデータにはノイズや曖昧性がほとんどないため，提案手法と近似モデルの結果に差がなかったと言える．

次に，mMLDA と近似モデルによって形成された人物概念を評価した．図 4.8（a），図 4.8（b），及び図 4.8（c）がそれぞれ，人手による分類，mMLDA の分類結果と近似モデルの分類結果を示している．他の概念と同様，図 4.8（a）を正解の分類として，両モデルによる分類結果を比較すると，それぞれ 75.75%及び 71.21%となった．このように mMLDA の学習では，知覚情報と概念間の関係を手がかりとして学習するため，下位層に生じる誤分類を概念間の関係によって修正することが可能である．

以上のように，下位概念の形成結果において，どの概念に対しても近似モデルに比べ mMLDA の方が，より正解に近い概念が形成された．これは，上位層を介

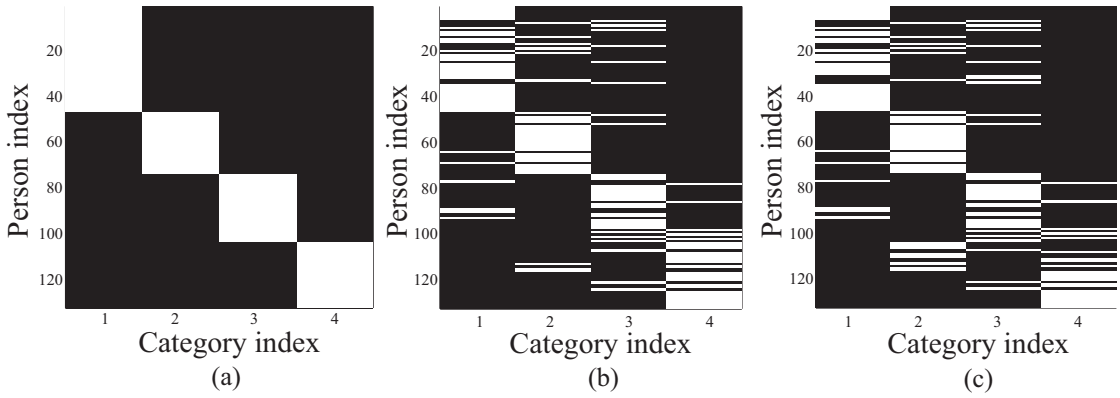

図 4.8: 人物の分類結果：(a) 正解, (b) mMLDA, (c) 近似モデル

して概念間の関係を手がかりとして用いた分類を行う mMLDA の方が，下位層の各概念に入力される知覚情報のみを用いる近似モデルに比べ，より人の感覚に近い分類が可能であることを意味する．

4.7.3 統合概念

mMLDA の上位層において形成された下位層の物体，動き，場所及び人物の関係を表現する上位概念に対して，表 4.3 にまとめた．表 4.3 は，上位層に形成された 30 個のカテゴリに対して，下位層の関係するカテゴリの組合せを示したもので，その中にいくつか意味のあるカテゴリが形成されていることが見て取れる．例えば，統合カテゴリ 21 において，人物カテゴリ「全員」，動きカテゴリ「口に運ぶ (3)」，場所カテゴリ「ダイニング」，3 つの物体カテゴリ「飲み物（缶）(17)」,「ペットボトル（18）」,「プラスチックカップ（25）」が一つのカテゴリに分類されている．これは,「人がダイニングで何かを飲む」という概念が上位に形成された結果であると考えられる．これと似たようなカテゴリとして,「人がダイニングで何かを食べる」が統合カテゴリ 25 として上位層に形成された．この統合概念は表 4.3 に示すように，統合カテゴリ 21 の人物，場所，動きカテゴリが同じであるが，それらと関係する物体カテゴリ（「茶碗（13)」,「カップヌードル（21)」,「野菜（玩具）(27)」）が異なっている．これに対して，統合カテゴリ 9 及び 22 は，動

表 4.3: mMLDA を用いた統合概念の形成結果

No	動き	物体	場所	人物
1	上下に振る	スプレー缶	庭	大人の男性
	塗る			
2	上に投げる	ぬいぐるみ	リビング	子供
		ボール		
3	中身を注ぐ	じょうろ	庭	大人の男性
4	上下に振る	ガラガラ	リビング	女の子
5	取り出す	ティッシュ箱	リビング	全員
		クッキー		
6	手に塗る	ハンドクリーム	リビング	大人の女性
7	皿を洗う	スポンジ	キッチン	大人の女性
		たわし		
8	中身を注ぐ	シャンプー	浴室	大人
9	左右に動かす	フローリングワイパー	ダイニング	大人の女性
10	取り出す	フローリングワイパー	ダイニング	大人の女性
11	上に投げる	マラカス	リビング	子供
	上下に振る			
12	履く	靴	玄関	全員
13	開ける	スナック	リビング	全員
14	包む	ラップ	ダイニング	大人の女性
15	持ち上げる	茶碗	ダイニング	全員
		カップヌードル		
		プラスチックカップ		
		飲み物（缶）		
		スプレー缶	庭	
16	置く	カップヌードル	ダイニング	大人
17	手に塗る	ハンドクリーム	リビング	女の子
18	中身をかける	ドレッシング	ダイニング	全員
		ソース		
		蜂蜜		
19	中身を注ぐ	ペットボトル	ダイニング	全員
		飲み物（缶）		
20	口に運ぶ	金属の食器	ダイニング	全員
21	口に運ぶ	ペットボトル	ダイニング	全員
		飲み物（缶）		
		プラスチック カップ		
22	左右に動かす	車（玩具）	リビング	男の子
23	積み重ねる	積み木	リビング	子供
	置く			
24	抱く	ぬいぐるみ	リビング	女の子
25	口に運ぶ	カップヌードル	ダイニング	全員
		野菜（玩具）		
		茶碗		
26	口に運ぶ	スナック	リビング	子供
27	置く	消臭剤	リビング	大人の女性
		除湿剤		
28	上下に振る	ドレッシング	ダイニング	全員
		ソース		
		ペットボトル		
29	すくう	ショベル	庭	大人の男性
30	ナイフで切る	野菜（玩具）	キッチン	大人の女性

きカテゴリ「左右に動かす（4)」が同じでも，それと関係する他の概念が異なるため，別のカテゴリとして分類された例である．統合カテゴリ 9 では，人物カテゴリが「大人の女性」と物体カテゴリが「フローリングワイパー（3)」と関係するため,「母がフローリングワイパーで掃除をする」という概念が形成されたと考えることができる．これに対して統合カテゴリ 22 では，人物カテゴリが「男の子（4)」，物体カテゴリが「車（玩具）(28)」と関係しているため,「男の子が車の玩具を走らせて遊ぶ」という概念が上位層に形成されていると言える．このように，同じ動きでも使用される物体や場所などが異なれば，意味が異なる上位カテゴリが形成されることが分かった．

他の例として，上位カテゴリ 3，8 及び 9 が挙げられるが，これらのカテゴリは同じ動きに対して，異なった場所や使用される物体が共起することで違うカテゴリとして分類されたと考えられる.「庭（5)」と「じょうろ（24)」の関係を表現する統合カテゴリ 3 は,「水遣りをする」という概念を意味するのに対し，統合カテゴリ 8 では,「浴室（6)」,「シャンプー（9)」と関係しているため,「シャワーを浴びる」が形成されていると言える．また，表 4.3 より「飲み物の中身を注ぐ」という概念は統合カテゴリ 19 に形成されていると考えることができる．一方，異なる動きのカテゴリ「積み重ねる（9)」と「置く（10)」が一つの上位カテゴリとして分類されている結果が統合カテゴリ 23 に現れている．このカテゴリは，物体カテゴリ「積み木（32)」，場所カテゴリ「リビング（2)」，人物カテゴリ「子供（3, 4)」と関係しており,「子供が積み木で遊ぶ」を意味する．以上のように，定性的には意味のある統合概念が形成できていると言えるが，統合概念は正解を定義することが難しいため，定量的に mMLDA と近似モデルを比較することができない．

そこでここでは前章と同様に，物体，動き，場所及び人物概念の関係を正確に表現できているかどうかを，同時確率で評価する．ここで，全ての下位概念 $\boldsymbol{z}^L = (z^O, z^M, z^P, z^U)$ の関係性は，その同時確率 $P(\boldsymbol{z}^L)$ で表現することができると考える．正解となる同時確率 $\hat{P}(\boldsymbol{z}^L)$ は，表 4.1 に示した各物体，動き，場所と人物の関係の学習サンプル数から，次式を用いて求めた．

$$\hat{P}(\boldsymbol{z}^L) = \frac{N_{\boldsymbol{z}^L}}{N} \tag{4.22}$$

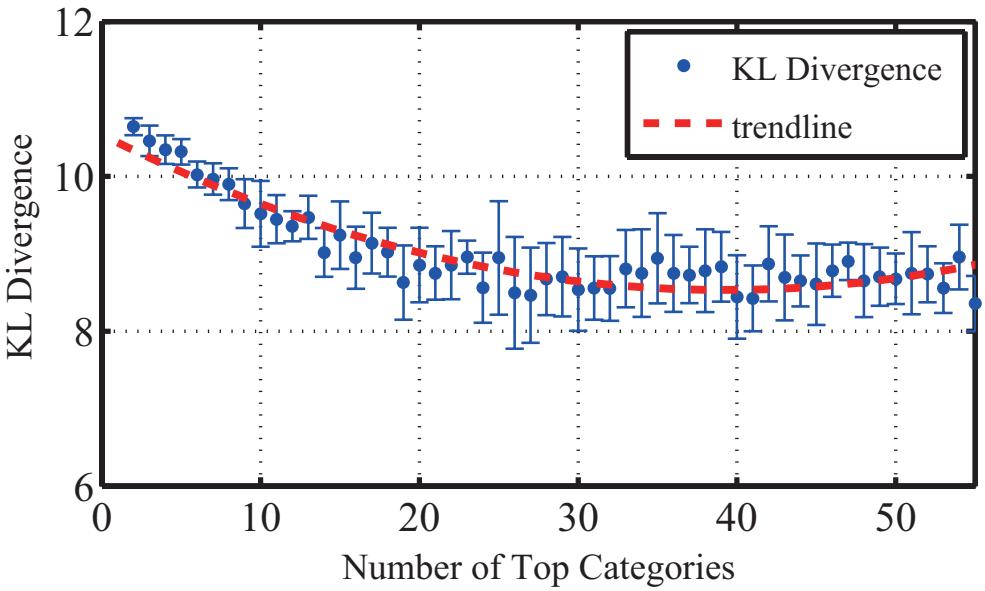

図 4.9: 上位カテゴリ数に対する同時確率分布の正解との KL ダイバージェンス

ただし，$N_{\boldsymbol{z}^L}$ は，下位概念 $\boldsymbol{z}^L$ の共起したデータ数であり，表 4.1 から求めることができる．また，N はデータの総数である．また，mMLDA と近似モデルで学習された同時確率 $P(\boldsymbol{z}^L)$ は，次のように計算可能である．

$$P(\boldsymbol{z}^L) = \sum_{z} P(z|\alpha) \prod_{z^C} P(z^C|z) \tag{4.23}$$

ここでは学習された同時確率 $P(\boldsymbol{z}^L)$ がどれだけ正解 $\hat{P}(\boldsymbol{z}^L)$ に近いかを，KL ダイバージェンスを用いて評価する．

$$D_{KL}\left(P(\boldsymbol{z}^L) \| \hat{P}(\boldsymbol{z}^L)\right) = \sum_{\boldsymbol{z}^L} P(\boldsymbol{z}^L) \log \frac{P(\boldsymbol{z}^L)}{\hat{P}(\boldsymbol{z}^L)} \tag{4.24}$$

近似モデルの結果と mMLDA の結果の正解との KL ダイバージェンスを求めた結果，それぞれ 11.34 と 8.53 となった．すなわち，mMLDA の方が近似モデルに比べ，より正確に概念間の関係を捉えられていると言える．

本実験では，MHDP を用いてカテゴリ数の決定を行った．上位カテゴリ数は 30

と推定されたが，カテゴリ数によって形成された上位カテゴリは変化してしまう．そこで，上位カテゴリ数の妥当性を評価するために，KL ダイバージェンスを用いて正解の同時確率と比較する．評価方法として前章と同じように，上位カテゴリ数を変化させて概念形成を行い $P(\boldsymbol{z}^L)$ を計算し，$\hat{P}(\boldsymbol{z}^L)$ との KL ダイバージェンスを計算した．その結果を図 4.9 にプロットする．図中の横と縦軸はそれぞれカテゴリ数と正解との KL ダイバージェンスを示している．カテゴリ数が少ない場合，KL ダイバージェンスが大きくなった．これは，少ないパラメータで概念間の関係を表現するため，正しく学習できないためであると考えられる．逆にカテゴリ数が大きい場合，多くのパラメータで表現できるため，正しくその関係を捉えることができ，正解との KL 距離が小さくなる．また，上位カテゴリ数がある一定以上大きくなると，KL ダイバージェンスは収束し変化しなくなるが，分類が細かくなってしまい概念が正しく形成できない可能性がある．実際，図 4.9 より，妥当な上位カテゴリ数は 30〜40 であることが見て取れる．従って，本実験において MHDP で推定された上位カテゴリ数 30 は適切であると言える．

4.7.4 未観測情報の予測実験

次に，未観測情報の予測性能を評価するために，観測した情報から未観測情報における概念の予測を行った．実験は以下の 4 つの場合を考慮して行った．

1. 物体の視・聴・触覚情報から，動き・場所・人物のカテゴリを予測
2. 動きの角度情報から，物体・場所・人物のカテゴリを予測
3. 場所の座標情報から，物体・動き・人物のカテゴリを予測
4. 人物の性別・年齢情報から，物体・動き・場所のカテゴリを予測

実験に用いたデータの組合せを，表 4.4 に示した．未観測情報の予測は mMLDA と近似モデルによって行い，それぞれの結果を比較した．予測結果の評価は，表 4.1 に基づいて，観測した情報に関係する全ての未観測概念のカテゴリを正解とする．例えば，観測した物体が「飲み物（缶）(17)」である場合，表 4.5 に示したカテ

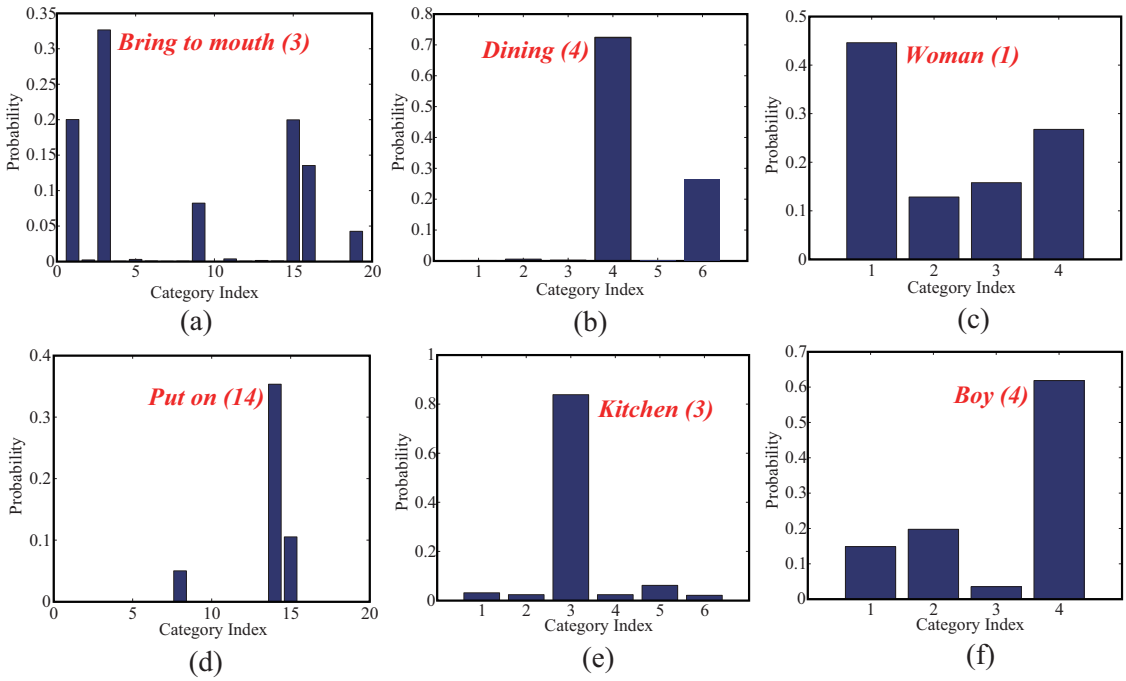

図 4.10: 「飲み物（缶）(17)」から mMLDA と近似モデルを用いた各概念のカテゴリの発生確率 :(a) mMLDA で動きカテゴリ,(b) mMLDA で場所カテゴリ,(c) mMLDA で人物カテゴリ,(d) 近似モデルで動きカテゴリ,(e) 近似モデルで場所カテゴリ,(f) 近似モデルで人物カテゴリ

ゴリを正解とした．mMLDA と近似モデルの予測結果を表 4.6 にそれぞれ示した．上記 4 つの場合において予測精度はどれも，mMLDA の方が近似モデルに比べ高い結果が得られた．これは，前節で述べたように，概念の形成において mMLDA の方が精度が高く，予測がし易いためである．

物体の情報から未観測情報を予測する実験において，図 4.3 に示した赤い枠で示した物体を認識用のデータとして用いて，残りの物体を学習用のデータとした．観測された物体のマルチモーダル情報（$\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h$）から動きカテゴリ z^M，場所カテゴリ z^P と人物カテゴリ z^U の予測を行った．図 4.10 は,「飲み物（缶）(17)」から予測された未観測である動きカテゴリ，場所カテゴリ，人物カテゴリが発生する確率 $P(z^M|\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h)$，$P(z^P|\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h)$ と $P(z^U|\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h)$ をそれぞれ表す．

表 4.4: 未観測情報のデータ

No	動き	物体	場所	人物
1	上下に振る	ガラガラ	リビング	女の子
2	上に投げる	ぬいぐるみ	リビング	女の子
3	左右に動かす	フローリングワイパー	ダイニング	大人の女性
4	皿を洗う	スポンジ	キッチン	大人の女性
5	皿を洗う	たわし	キッチン	大人の女性
6	手に塗る	ハンドクリーム	リビング	大人の女性
7	テーブルに置く	消臭剤	リビング	大人の女性
8	テーブルに置く	除湿剤	リビング	大人の女性
9	中身を注ぐ	シャンプー	浴室	大人の男性
10	取り出す	ティッシュ箱	リビング	大人の男性
11	包む	ラップ	ダイニング	大人の女性
12	持ち上げる	茶碗	ダイニング	大人の男性
13	上下に振る	ドレッシング	ダイニング	大人の男性
14	中身をかける	蜂蜜	ダイニング	男の子
15	上下に振る	ソース	ダイニング	男の子
16	持ち上げる	飲み物（缶）	ダイニング	男の子
17	口に運ぶ	ペットボトル	ダイニング	大人の女性
18	口に運ぶ	スナック	リビング	男の子
19	持ち上げる	カップヌードル	ダイニング	大人の男性
20	開ける	スナック	リビング	大人の男性
21	持ち上げる	スプレー缶	リビング	女の子
22	中身を注ぐ	じょうろ	庭	大人の男性
23	持ち上げる	プラスチックカップ	ダイニング	大人の女性
24	すくう	ショベル	庭	大人の男性
25	口に運ぶ	野菜（玩具）	ダイニング	男の子
26	左右に動かす	車（玩具）	リビング	男の子
27	上に投げる	マラカス	リビング	男の子
28	履く	靴	玄関	大人の男性
29	上に投げる	ボール	リビング	男の子
30	積み重ねる	積み木	リビング	男の子

表 4.5: 飲み物（缶）に関係する物体，場所，人物のカテゴリ（カッコ内の数字はカテゴリ番号）

動き	物体	場所	人物
持ち上げる（1）	飲み物（缶）（17）	ダイニング（4）	女の子（3）
口に運ぶ（3）	飲み物（缶）（17）	ダイニング（4）	大人の男性（2）
口に運ぶ（3）	飲み物（缶）（17）	ダイニング（4）	女の子（3）
上下に振る（6）	飲み物（缶）（17）	ダイニング（4）	大人の女性（1）
上下に振る（6）	飲み物（缶）（17）	ダイニング（4）	大人の男性（2）
中身を注ぐ（15）	飲み物（缶）（17）	ダイニング（4）	女の子（3）
中身を注ぐ（15）	飲み物（缶）（17）	ダイニング（4）	男の子（4）

mMLDA を用いた動きカテゴリの予測結果（図 4.10（a））において，正しく「持ち上げる（1）」や「口に運ぶ（3）」といった動き（表 4.5 を参照されたい）を予測することができているが，近似モデルを用いた予測の結果（図 4.10（d））では，「中身をかける（14）」といった動きが高い確率で予測されている．これは，近似

表 4.6: 未観測情報の予測精度

観測した情報	物体	動き	場所	人物
mMLDA				
視・聴・触覚	-	**76.67%**	**80.00%**	**73.33%**
角度	**86.67%**	-	**80.00%**	**90.00%**
座標	**76.67%**	76.67%	-	**100%**
性別・年齢	**80.00%**	**83.33%**	**86.67%**	-
近似モデル				
視・聴・触覚	-	66.67%	70.00%	70.00%
角度	76.67%	-	73.33%	80.00%
座標	70.00%	76.67%	-	90.00%
性別・年齢	76.67%	73.33%	80.00%	-

モデルの分類結果では，物体の「飲み物（缶）(17)」と「ドレッシング（14)」が同じカテゴリに分類されてしまったため,「ドレッシング（14)」に関係する「中身をかける（14)」が予測されてしまったと考えられる．このように，近似モデルでは物体や動きなど下位層の概念を独立に学習するため，下位層における分類の誤りを概念間の関係を通して修正することができず，予測精度の低下を引き起こしたと考えられる．

また，mMLDA を用いた場所カテゴリの予測結果（図 4.10（b））において，正しく「ダイニング（4)」を予測することができているが，近似モデルの結果（図 4.10（e））では，誤った場所カテゴリである「キッチン（3)」が最も高い確率で予測されている．しかし，人物カテゴリの予測結果では，表 4.5 に示した通り，全ての人物カテゴリに関係するため，mMLDA と近似モデルの予測結果（図 4.10（c）と図 4.10（f））が異なったとしても，どれも正解となるため，両方正しく予測できている． 以上のように，近似モデルに比べ mMLDA の予測性能が高いことが分かる．

4.7.5 単語予測実験

本実験では，入力されたマルチモーダル情報に対して単語の予測を行った．まず単語情報に対する概念選択の結果について説明する．概念の選択を行うために，前節で述べたように相互情報量を計算しそれを重みとして用いる．ここでは，学習データの単語情報について，各概念クラスに対する重みを求め，その結果を図

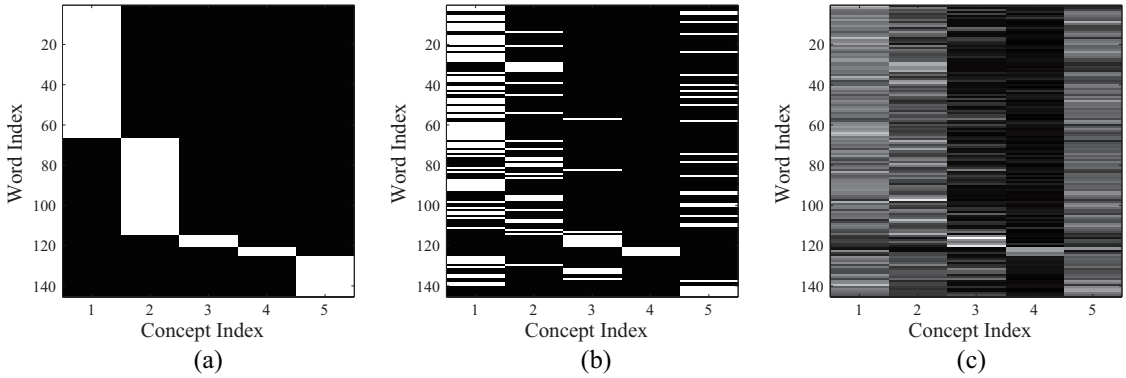

図 4.11: 概念選択の結果

4.11 に示した． 図 4.11 の横と縦軸はそれぞれ各概念の番号（1 から 5 まで順番に，物体，動き，場所，人物と，統合概念を表す）と単語のインデックスを表す．図 4.11（a）は各単語に対して人が定義した正解となる概念クラスを示す．mMLDA の学習結果から求めた各単語と概念の相互情報量を図 4.11（b）と（c）にプロットする．図 4.11（b）は，ある単語において全ての概念クラスに対して計算した相互情報量の中から最大となる概念クラスを表しており，各概念の相互情報量を図 4.11（c）に示した．単語における概念選択を評価するために，予め人が用意した単語と概念の対応リスト（表 4.7）を用いて正解率を求めたところ，68.75%の正解率であった．また，各概念クラスに対する正解率の詳細を表 4.8 に示した．概念選択の精度を計算するために，相互情報量の最大となる概念クラスを採用し比較を行った． 精度としての結果は，まだ向上する必要がある．しかし実際の単語予測において，相互情報量は重みとして用いるため，最大値による概念選択に誤りが生じたとしても，単語発生確率の結果と合わせることで，正しい単語が予測されるケースが多い．特に相互情報量の結果において，正解となる概念がわずかな差で 2 位となる場合は，かなりの確率で正しい単語予測を行うことができる．これより，相互情報量を用いた概念選択は単語予測の重みとして，十分な精度であると考えることができる．

次に，単語予測実験について説明する．まずは，物体概念から単語の予測を行った．結果の一例として，図 4.12 に示す物体の「ぬいぐるみ」から予測された単語

表 4.7: 各概念を表現する単語の一部

物体	動き	場所	人物	統合
ガラガラ	かける	キッチン	女の子	塗料
スナック	運ぶ	ダイニング	男の子	飲む
飲み物	塗る	リビング	父	食べる
ペットボトル	動かす	玄関	母	拭く
ぬいぐるみ	投げる	庭		遊ぶ

表 4.8: 各概念における概念選択の正解率

	物体	動き	場所	人物	統合	全概念
単語数	91	48	6	4	32	181
正解率	78.78%	53.33%	100%	100%	56.52%	68.75%

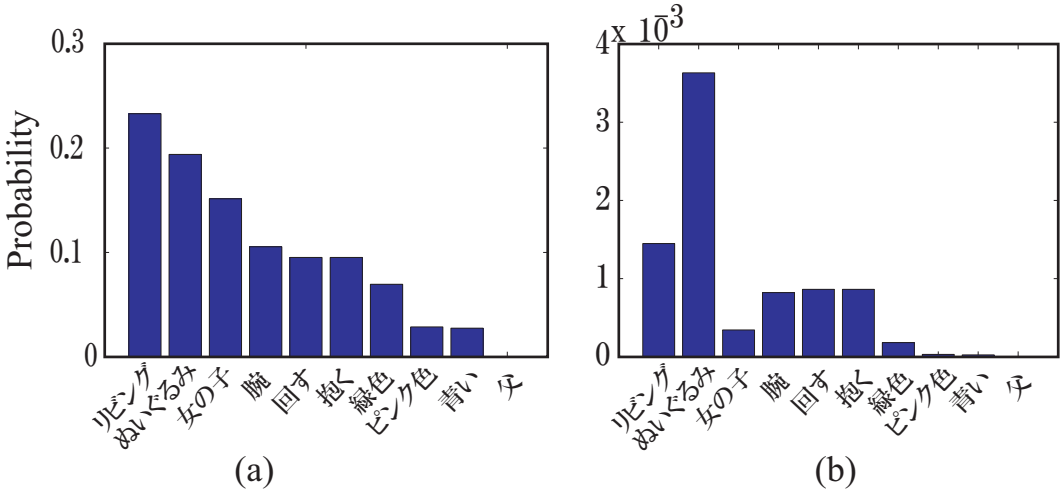

図 4.12: 「ぬいぐるみ」からの単語予測 :（a）単語の発生確率,（b）相互情報量による重み付けをした単語発生確率

について述べる．図 4.12（a）は「ぬいぐるみ」の視・聴・触覚情報が観測されたときの単語の発生確率を表し，これより統合概念を表す「リビング」という単語が一番高い確率で予測されていることが分かる．一方，相互情報量を各概念に対する重みとして計算し単語発生確率にかけた結果が図 4.12（b）である．これより，

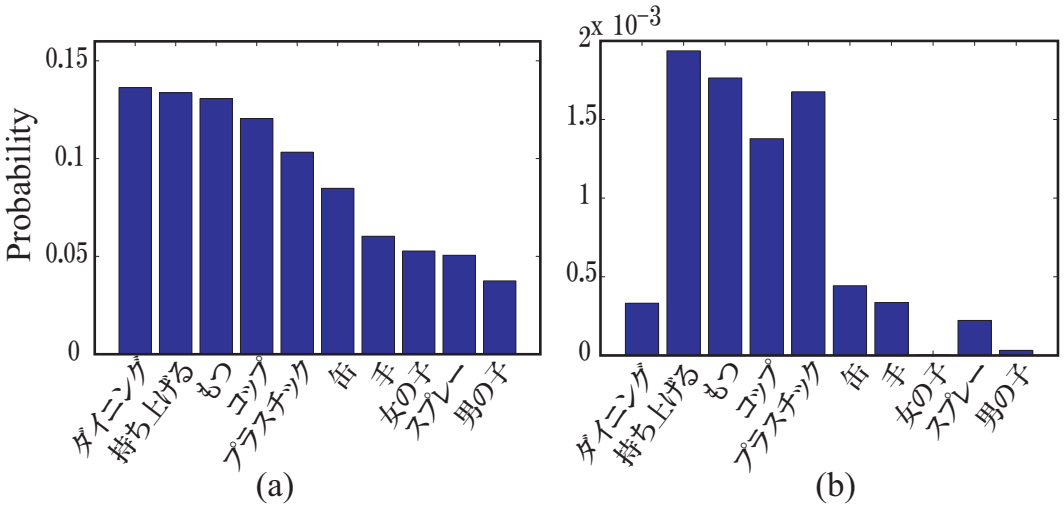

図 4.13: 「持ち上げる」からの単語予測：(a) 単語の発生確率, (b) 相互情報量による重み付けをした単語発生確率

物体概念を表す「ぬいぐるみ」の単語が一番高い確率で予測されるようになったことが分かる．他の例として,「スプレー缶」から予測された単語の発生確率において,「庭」という単語が最も高い確率で予測されたが，この予測結果に単語の相互情報量による重みを付けると,「スプレー」と「缶」という単語が正しく予測されるようになった．このように，相互情報量の重み付けによって，単語を正しく予測することが可能である．

同様に，動き情報のみが観測されたときの単語予測において,「持ち上げる」の動き情報から単語の予測を行った結果を図 4.13 に示した．図 4.13（a）から「ダイニング」といった単語が高い確率で予測された．一方，図 4.13（b）の結果から，動き概念以外に関係する単語の確率は，相互情報量の重み付けによって低くなり,「持ち上げる」や「もつ」といった単語が高く予測されるようになった．しかし，今回の学習データにおいて,「持ち上げる」という単語は統合概念を表す単語と設定したにも関わらず，相互情報量の重み付けにおいても，動き概念と統合概念との相互情報量の値がほぼ同じとなったため,「持ち上げる」という動き情報に対して，動き概念を表す「もつ」が 2 番目に高く予測される結果となった．他の例として,「口に運ぶ」という動きに対する単語の予測では,「ダイニング」という単語

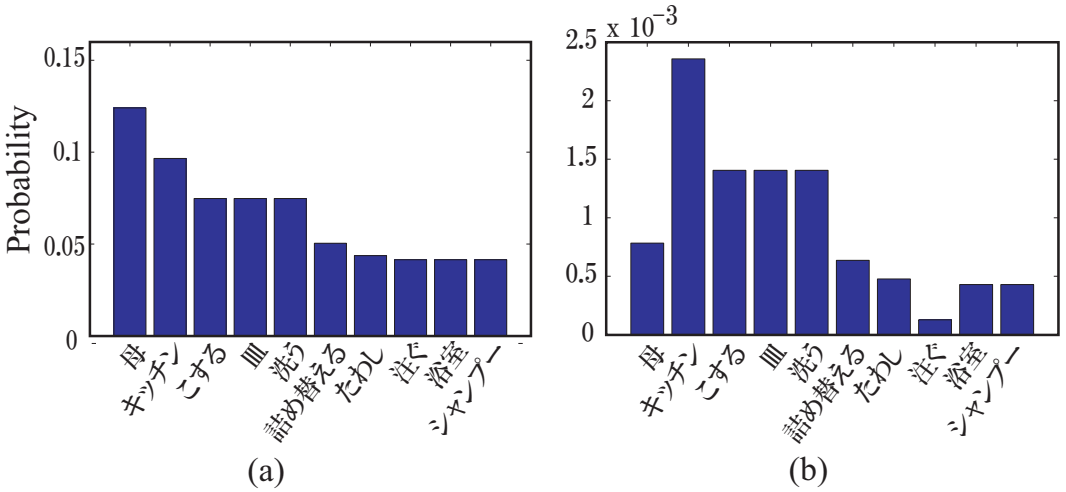

図 4.14: 「キッチン」からの単語予測：(a) 単語の発生確率, (b) 相互情報量による重み付けをした単語発生確率

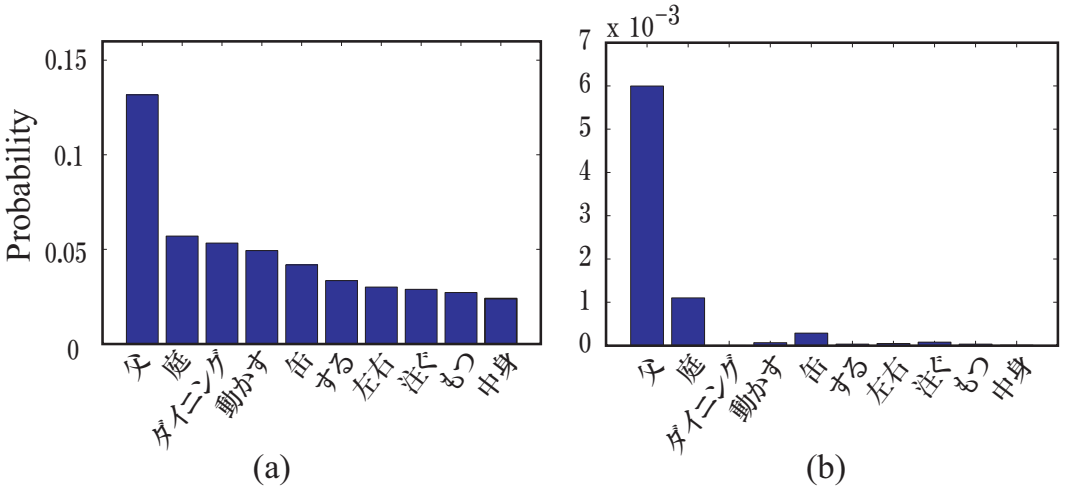

図 4.15: 「大人の男性」からの単語予測：(a) 単語の発生確率, (b) 相互情報量による重み付けをした単語発生確率

が最も高く予測されたが，この予測結果に相互情報量による重みを付けると,「口」や「運ぶ」といった正しい単語が予測される結果となった.

図 4.14 に示した「キッチン」の場所情報から予測された単語の結果も，相互情

報量を重み付けとして用いた提案手法の有効性を示している．単語発生確率（図 4.14（a））の結果において,「母」という単語が高く予測されたが，提案手法を用いた結果（図 4.14（b））では，正しく「キッチン」といった場所概念に関係する単語が予測された．また人物の予測結果では，本実験に用いたデータにおいて，単語発生確率と提案手法はそれぞれ正しい単語を示した（図 4.15）．

以上の結果より，多様な概念において教示文に含まれる単語には，どの概念に結び付けるかという情報がないため，学習したモデルを用いて単語の発生確率をそのまま単語予測の結果として扱うと，その概念に関係しない単語が多く発生する結果となる．この問題を解決するために，単語と概念との相互情報量を手がかりとして単語発生確率に重み付けする単語予測手法が予測性能を大きく向上することが分かった．

4.7.6 観測情報からの言語生成

ここでは，提案する文生成の有効性を検証するために，表 4.1 のデータを用いて実験を行った．まず，提案手法によって獲得した文法と，人がラベル付けした正解文法を図 4.16 に示す．これにより，提案手法を用いることで，人手に近い文法が獲得できることが分かる．この文法を用いて各データに対して，前節で記述した「Method 1」及び「Method 2」を用いて文生成を行った．文例を以下に示す．

S1: 母 ダイニング 茶碗 手 もつ 持ち上げる
M1: 母 ダイニング 持ち上げる 父
M2: 母 ダイニング 茶碗 手 もつ 持ち上げる

S2: 父 ダイニング 黒い 茶碗 手 もつ 持ち上げる
M1: 父 ダイニング 持ち上げる 持ち上げる 父 持ち上げる 父
M2: 父 ダイニング 茶碗 手 もつ 持ち上げる

S3: 母 ダイニング カップ ヌードル もつ 持ち上げる

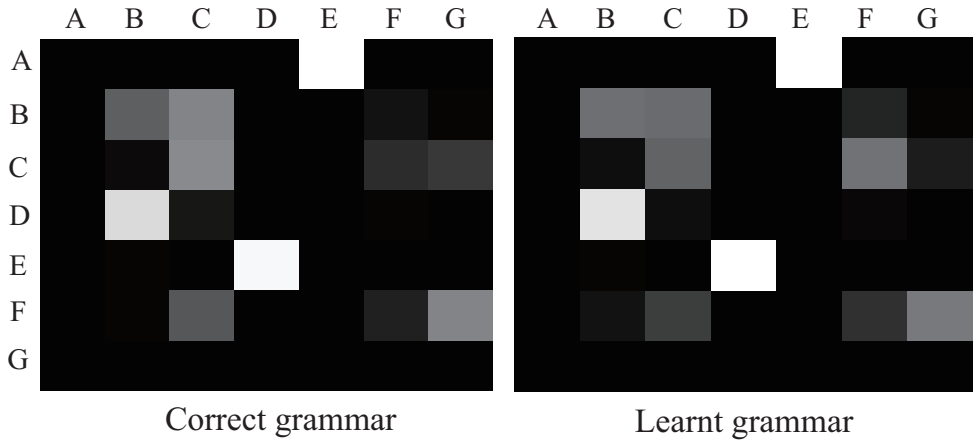

図 4.16: 獲得した文法と正解文法：図中の A，B，C，D，E，F，G はそれぞれ BOS，物体概念，動き概念，場所概念，人物概念，統合概念，EOS を表している

M1: 母 ダイニング 男の子 父
M2: 母 ダイニング カップ ヌードル もつ 持ち上げる

S4: 女の子 リビング ハンド クリーム 手 つける 塗る
M1: 振る リビング クリーム クリーム 塗る
M2: 女の子 リビング ハンド クリーム つける 塗る

S5: 母 ダイニング 透明 ラップ 何 覆う 包む
M1: 母 ダイニング ラップ ラップ 母 母 男の子
M2: 母 ダイニング 透明 ラップ 使う 何 覆う 包む

ただし，M1 と M2 はそれぞれ,「Method 1」と「Method 2」を表しており，S1，S2，S3，S4，S5 はコーパス内の文である．このように，概念遷移のみで生成された文（M1）は，同じ単語の繰り返しが多く見られた．また，各概念に対して単語の予測が間違ってしまうと S5 のように異なる単語で文が生成されてしまう．これに対して，言語モデルを考慮した M2 を用いるとよりコーパスに近い文が生成されることが分かる．

文生成の定量的な評価として，生成された文に対して，教示発話と比較して，単語 2-gram の BLEU-2 スコア，及び単語 3-gram の BLEU-3 スコア [65] を計算した．BLEU は機械翻訳システムの自動評価として主に利用されている評価基準であり，機械翻訳結果と参照訳との n-gram のマッチ率に基いている．一般的に BLEU スコアは，翻訳文と参照訳の 1-gram から n-gram について幾何平均を計算するスコアと文の長さを考慮するスコア BP（Brevity Penalty）の掛けあわせで算出される．n-gram は文中の単語と単語同士の順番を表現するため，単語の正しさ（1-gram）及び流暢さを表すことができる．一方，BP では文の長さのペナルティとなっており，短い文ほどペナルティが大きく，その値が 0〜1 の範囲となる．従って，BLEU スコアは 0〜1 の値として計算され，値が高ければ良い文であると考えることができる． 全データに対して,「Method 1」及び「Method 2」の BLEU-2 スコアの平均はそれぞれ 0.28 と 0.61 となり，BLEU-3 スコアの平均はそれぞれ 0.16 と 0.45 となった．これより,「Method 1」に比べ「Method-2」が生成する文の文らしさを，大きく向上することができていることが客観的にも分かる．

以上の結果を踏まえて考察する．観測情報からの文生成では，学習した文法からの概念列をサンプリングするという処理から始まる．この段階において，一つのサンプルが正しい概念列となる可能性は，複数のサンプルよりも当然低い．言い換えると，複数のサンプルを用意することで，正しい概念列を持つ文が作成できる可能性が高くなる．この点において,「Method 1」のように一つのサンプルしか用いない手法では，概念列の選択ミスが生じると，正しくない文が生成されてしまう．これに対して「Method 2」では，選択ミスが生じたとしても，複数のサンプルを用いるためその中からサンプルを選ぶことで問題を回避することができるが，複数のサンプルをどのように選択すればいいかという問題を考えなければならない．また，各サンプルに対してどのような単語を配置すれば良い文が作れるかという問題も考慮する必要がある．単語の選択については，概念と単語の結び付けより解決することができるが，前節に述べたように単語における概念選択の結果は万全ではないため，確率的に全ての可能性を考慮すべきである．この点に関して,「Method 1」ではその概念に最も関係する単語しか考慮しないため，予測に誤りが生じてしまうと修正することができず，おかしいな文が生成されてし

まう．また上述のように，文として考えるときの単語同士の関係を表現する言語モデル（n-gram）も文の自然さの観点から重要な要素となる．このように，文として考える際に各概念に対する単語を一つ選択するのではなく，複数候補を用意し文全体のつながりを言語モデルで評価すれば，良い文のサンプルが用意できると考える．最後の問題は，複数の文からの候補の選択であるが，提案手法では文の長さと生成確率を考慮した文の選択を行っている．従って，文の生成において，複数のサンプルを用意しその中から最も高い生成確率を選択する「Method 2」は，一つのサンプルで単語同士のつながりを考慮しない「Method 1」の性能を大幅に向上することができる．

4.8 まとめ

本章ではまず，前章で提案した mMLDA を拡張し，より多様な概念を獲得する手法を提案した．提案モデルにおいては，物体，動き，場所，人物概念が下位概念として下位層において表現され，それらの関係性としての上位概念が上位層に表現される．こうして形成された多様な概念を利用することで，様々な予測を行うことが可能である．実験結果より，提案した mMLDA が近似モデルに比べ，高い予測性能を示すことが明らかとなった．これは，上位・下位概念が相互に影響し合うことが，階層的な概念形成において重要であることを意味する．

本章ではさらに，概念と単語の結び付け手法を提案した．多様な概念を扱っている mMLDA において，どの単語がどの概念と結び付くべきかを考える必要があり，ここではこの問題を相互情報量により解決し，単語予測実験を通して相互情報量を用いない場合と比べより高い予測結果を示すことを明らかにした．また，教示文内の単語が指す概念の遷移に基づく文法の学習を可能な手法を提案した．こうした多様な概念，概念間の関係，語意，及び文法を学習することで，観測情報からの文生成を行うことを可能とした．

提案した mMLDA は，人の行動に含まれている物体や動きのみならず場所や言語（単語）の共起性を手がかりとして学習を行ったことに相当する．提案モデルは，2 章で述べた確率的な知識表現となっており，人の行動を観測することでロ

ボット自身が知識を獲得することを可能にする．しかし mMLDA は，人の活動における行動と行動の時間的な関係を表すことができていない．次章では人の行動文脈として，人の行動における共起性だけではなく，行動と行動の関係を表すモデルを提案する．さらに，行動文脈だけではなく音声命令など様々な文脈を統合し，ロボットが行動決定する手法について議論する．

第5章　動作概念と文脈の統合によるロボットの行動決定

5.1　はじめに

本論文では，ロボットが経験した事物の階層的なカテゴリ分類を通して多様な概念を形成し，それらの概念を利用して様々な予測を行うことで事物に対する真の理解を実現することを目的としている．これは前章で議論した多層マルチモーダルLDA（mMLDA）によって実現することが可能であることを示した．この枠組によってロボットは，観測する事物を理解し，理解したことを言語として出力することが可能となる．

一方，実際のコミュニケーションは，背景知識や周辺の状況などといった文脈を考慮しなければ成立しない．つまり，事物に対する理解をより柔軟に行うためには，学んできた多様な概念を活用した上で，様々な文脈を考慮する必要がある．本章では，ロボットが人と生活する上で，様々な文脈においてどのように行動決定するかを議論する．

一般に，ロボットはユーザの命令に応じて行動する．ユーザの命令は一つの行動に対して様々な形で存在し得るため，ロボットは命令を解釈して適切な行動を決定しなければならない．また音声による命令では，音声認識誤りが生じる可能性がある．こうした状況の中で柔軟に対応できることが，人々の生活を支援するロボットに要求される．

ロボットがユーザの命令を正しく解釈するための手がかりとして，命令を受けた際の文脈が考えられる．例えば，ユーザが普段ソファーでテレビを見ているときに，お菓子を食べながらお茶を飲んでいるということを知っていれば，ユーザが「お菓子を持ってきて」とロボットに命令した際の音声認識に誤りが生じたと

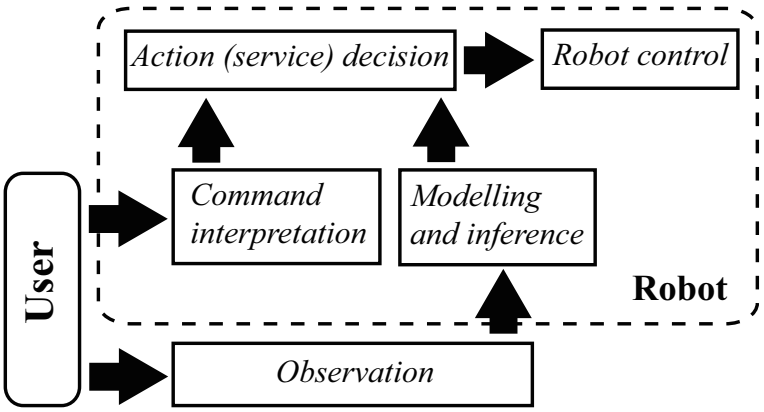

図 5.1: 提案手法の概要

しても，そのときにソファーでテレビを見ていてお茶を飲んでいるという文脈を用いれば，ロボットが適切に判断をして正しい行動をとることができるかもしれない．また，ユーザが日々行っている行動をロボットが学習できれば，ユーザからの命令がなくても，ユーザの次の行動をロボットが予測し，適切なサービスの提供が実現できると考えられる．

本章ではそのようなシナリオを想定し，ロボットがユーザの生活に密着し行動を観測することで，行動パターンを学習することを考える．この学習した行動パターンを，次の行動の予測に利用する．ロボットは，この予測した次の行動を文脈として音声命令や場所などと統合することで最終的なサービス行動を決定する．図 5.1 に全体的な考え方を示す．本章で考える重要な問題は，1) ユーザの観測，2) モデル化・予測，3) サービス・行動決定である．1) に関しては，人を常に観測し続けることについての技術的な困難さがある．スマートホームのような環境を考えることもできるが，そうでない環境においてもロボットがなるべく人をセンシングできれば有用であろう．本章ではこの問題に対して，センシングを考慮した人追跡手法を用いることを考えており，そのための基礎的な検討を行う．しかし実際のロボットに実装するためには，ロボットの制御などが必要となる．本章では，1）の実装より，2）の問題に焦点を当てるため，1）の問題に関しては後の議論とする．2) は，どのように行動をモデル化・予測するのかということである．これについても様々考えられるが，ここでは人の動きとその際に使う物体の関係性を

教師なしで学習し予測することを試みる．従って重要なのは，動作の学習と動作と物体の結び付きの学習である．動作時系列の学習には階層ディリクレ過程隠れマルコフモデル（HDP-HMM）[66] を，物体と行動の関係性の学習には前章で提案した多層マルチモーダル Latent Dirichlet Allocation（mMLDA）を用いる．また，mMLDA によって記号化された行動を n-gram（行動言語モデル）で表現し予測を行う．さらに，3) については様々考えられるが，ここでは「何かを持ってくる」というサービスを実現する．つまり，ロボットは人の行動パターンから次を予測し，必要となるであろうものを持ってくるというサービスを行うが，人からの「〜持ってきて」といった命令があった場合は，その解釈と予測を統合することでより精度よくサービスが実行できるのではないかと考える．

関連研究として，人の行動学習・認識や，物体と動きの関係性の学習などが挙げられる [32, 67–69]．文献 [67, 68] では，隠れマルコフモデルや Support Vector Machine（SVM）を用いた人の行動認識及び意図推定を行っている．しかし，人の行動に関係する物体の関係性は考慮されていない．また，教師あり学習に基く動きと物体の関係性の学習が [32, 69] において研究されているが，画像内の物体認識精度の向上が主眼となっている．これらに対して本研究では，人の行動を教師なしで学習・認識し，音声命令や場所などの文脈と統合し，適したサービスを決定する問題を扱う．

5.2 提案手法

5.2.1 提案手法の概要

図 5.2 に提案手法の全体像を示す．本章では，ロボットは家庭でユーザと共に暮らしていることを想定し，人の音声命令を聞きながらその人の動作と，動作を行っている際に関係している物体及び位置を観測する．そして，ユーザの行動パターンを教師なしで学習するのであるが，その際にまず使っている物体を認識しトラッキングすることで，動作の分節化を行う．これは，同じ物体を使っている間が一つの行動としての塊であると仮定し，時系列パターンを区切ることを意味する．そのように区切った動作時系列（関節角の情報）と物体の関係性は，mMLDA

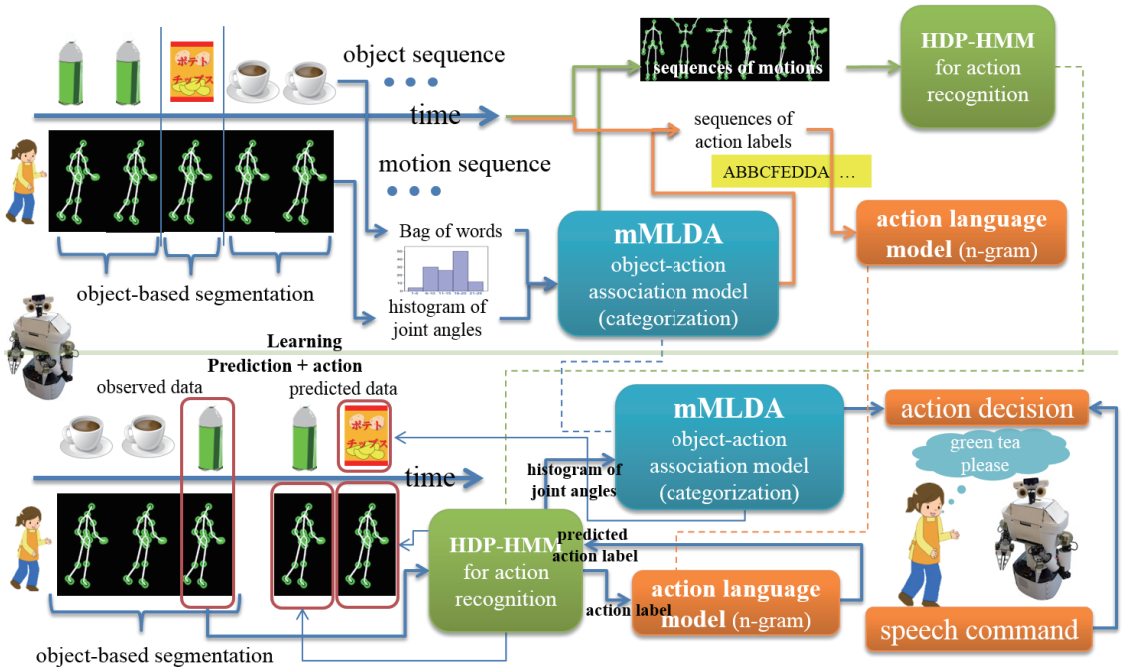

図 5.2: 提案手法の全体像

によってモデル化（カテゴリ分類も含む）することができる．つまり現在の行動を認識し，その後に起こる行動を予測できれば，mMLDA によって確率的にユーザが使うであろう物体を予測し，それを持ってくるというサービスを実現可能である．これを「動作－物体関係モデル」と呼ぶ．現在の動作の認識は，分節化された動作の時系列を HDP-HMM を用いてモデル化することで実現する．これを「動作認識モデル」と呼ぶ．また行動の予測については，学習データから行動の n-gram である,「行動言語モデル」を計算することで実現する．こうした動作認識モデルや行動言語モデルは，動作を分節化し，mMLDA を用いたカテゴリ分類に基づく記号化によって実現されることに注意されたい．

5.2.2 ロボットによる能動的センシング

ここでは前述した 1) ユーザの観測の問題に対する解決策として，ロボットによる能動的センシングについて議論する．ロボットがユーザをよりよく観測し続け

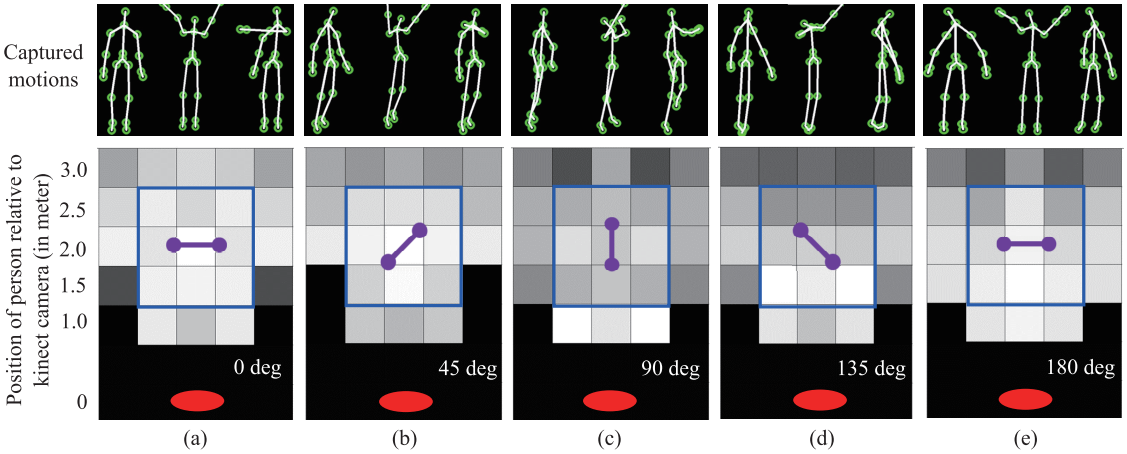

図 5.3: KINECT より取得された骨格情報のスコアマップ

るためには，追跡するだけではなく，関節角のセンシングに有利な位置に移動することが重要である．本章では，頭部に KINECT と台車に LRF が搭載されたロボットを用いることを前提に，ユーザの姿勢推定に有利な位置を考慮しながら移動することを考える．

本章では，ユーザの姿勢推定には KINECT を用い，ロボットの自己位置推定には LRF を用いる．KINECT は，ゲームコントローラとして開発されたものであるため，カメラから $2m$ 程度離れ正面に向かって動作を行うことが想定されている．しかし普段我々は，必ずしもロボットの前で行動するとは限らない．そこで，実際 KINECT でユーザの姿勢推定が可能な相対位置関係を検証し，ロボットがその制約をなるべく満たすようにユーザの追跡を行うこととする．実際，人の姿勢が取得可能な領域を調べるための予備実験を行った．

本実験において，KINECT をある位置 $\boldsymbol{x}_0 = (x_0, y_0, \Omega_0)$ に固定し，KINECT から領域 $D(x, y)(x_0 - \frac{\ell_x}{2} \leq x \leq x_0 + \frac{\ell_x}{2}, y_0 \leq y \leq \ell_y)$ をグリッド化した．各グリッド $R(x_{gx}, y_{gy})(gx \in [1, G_x], gy \in [1, G_y])$ に対して，3 種類の動作をそれぞれ行った．ただし，各動作において，KINECT に対する角度 $\Omega_t(t \in [1, T])$ を T 種類設定した．本章では，ℓ_x，ℓ_y，G_x，G_y，T の値をそれぞれ，$2.5m$，$3.0m$，5，6，5 に設

定した．そして，以下の式より各グリッド R_g のスコア $S(R_g)$ を求める．

$$S(R(x_{gx}, y_{gy})) = \frac{\sum_{f=1}^{F_{gxy}} \sum_{n=1}^{N_{\dot{\psi}}} \delta(\dot{\psi_{fn}})}{F_{gxy}}, \tag{5.1}$$

ただし，F_{gxy}，$N_{\dot{\psi}}$ はそれぞれグリッド $R(x_{gx}, y_{gy})$ における動作のデータ数と角速度 $\dot{\psi}$ の次元数を表す．また，ψ_{fn} は $\dot{\psi}$ の f 番目の n 次元目の要素であり，以下のように求める．

$$\delta(\dot{\psi_{fn}}) = \begin{cases} 1 & (\psi_{fn} \geq \text{閾値}) \\ 0 & (\psi_{fn} < \text{閾値}) \end{cases} \tag{5.2}$$

式（5.1）を用いて姿勢推定可能な領域のマップを計算した結果を，図 5.3 に示す．図中のグリッドマップは $3.0m \times 2.5m$ の部屋で計算しグリッドサイズを $0.5m$ とした．人がある姿勢（図中の紫棒であり，KINECT に対する角度は白い数字で表す）において，KINECT（図中の赤楕円）の前に立って動作を行うとき，骨格情報がどれだけ正確に取得できるかというスコアマップを図 5.3（a）〜（e）に示す．各グリッドにおいて，スコアが高いほど白く描画されている．

図 5.3 より，青い矩形に囲まれたグリッド内で動作を行うとき，姿勢推定の精度が上がることが分かった．また，当然ながら人が KINECT に対して，正面にあるときのスコアが高かった．本章ではこれらの事実をもとに，人が領域内にあるようにロボットが人の位置を計算し，人に対して正面にあるように移動することで，人を観測し続ける．

5.2.3 問題設定

本章で実現したいサービスは，予測した行動に必要となる物体をユーザに先回りして届けることである．その際，ユーザがどこで動作を行うかという場所文脈や命令された音声文脈などを考慮すればより正確なサービスが行えるであろう．この問題設定は，全ての情報が与えられた際，持ってくるべき物体を推定する問題に置き換えることができる．すなわち，現在の時刻 $t-1$ にロボットが観測したユーザ

の動き $\boldsymbol{m}^{(t-1)}$，物体 $o^{(t-1)}$，位置 x，音声 S に対して以下の問題を解くことになる．

$$\hat{o}^{(t)} = \underset{o^{(t)}}{\operatorname{argmax}} P(o^{(t)} \big| \boldsymbol{m}^{(t-1)}, o^{(t-1)}, x, S) \tag{5.3}$$

上式を直接的に計算するのは困難であるため，次のように近似する．

$$\hat{o}^{(t)} = \underset{o^{(t)}}{\operatorname{argmax}} P(o^{(t)} \big| \boldsymbol{m}^{(t-1)}, o^{(t-1)})^{\omega_1} P(o^{(t)} \big| x)^{\omega_2} P(o^{(t)} \big| S)^{\omega_3} \tag{5.4}$$

ただし，$P(o^{(t)} \big| \boldsymbol{m}^{(t-1)}, o^{(t-1)})$，$P(o^{(t)} \big| x)$，$P(o^{(t)} \big| S)$ はそれぞれ，行動文脈，場所文脈及び音声命令を表しており，ω_1，ω_2，ω_3 は各文脈に対する重みである．

各文脈の重みの決め方は様々な手法が存在する．例えば，重みのアクティブな学習 [71] を考えることもできるが，ここでは SVM による学習を用いる．具体的には，次節で説明する各文脈 $C \in \{C_1 = \text{行動}, C_2 = \text{場所}, C_3 = \text{音声}\}$ より予測される物体の確率分布 $\boldsymbol{P}^C = (p_1^C, p_2^C, \cdots, p_O^C)$ を一つのヒストグラム $\boldsymbol{h}^C = (\boldsymbol{P}^{C_1}, \boldsymbol{P}^{C_2}, \boldsymbol{P}^{C_3})$ として，SVM の入力データとする．ただし，O は物体カテゴリ数を表す．学習フェーズにおいて，入力データ $\boldsymbol{h}^C$ と正解となる物体カテゴリ o^C の組を用意し，SVM [48] を用いて学習する．認識フェーズでは，与えられた入力データ $\boldsymbol{h}_{\text{in}}^C$ に対して，SVM で学習したモデルを用いて認識する．

5.3 行動文脈

行動文脈 $P(o^{(t)} \big| \boldsymbol{m}^{(t-1)}, o^{(t-1)})$ は，現在の行動から次の予測された行動に対する物体の関係性を表す．ここで，現在の時刻 $t-1$ に観測したユーザの動き $\boldsymbol{m}^{(t-1)}$，物体 $o^{(t-1)}$ が与えられた場合，行動文脈を以下のように定義する．

$$P(o^{(t)} \big| \boldsymbol{m}^{(t-1)}, o^{(t-1)}) = \sum_{a_i^{(t)}, a_j^{(t-1)}} P(o^{(t)}, a_i^{(t)}, a_j^{(t-1)} \big| \boldsymbol{m}^{(t-1)}, o^{(t-1)}) \tag{5.5}$$

ただし，$a_i^{(t)}$，$a_j^{(t-1)} (a_* \in \boldsymbol{A}^{\text{K}} = \{a_1, a_2, \cdots, a_K\})$ はそれぞれサイズ K の行動集合 $\boldsymbol{A}^{\text{K}}$ に対して，時刻 $t-1$ と時刻 t における行動を表している．また，チェインルールと独立性を用いて，$P(o^{(t)}, a_i^{(t)}, a_j^{(t-1)} \big| \boldsymbol{m}^{(t-1)}, o^{(t-1)})$ を以下のように書き表すこ

とができる．

$$P(o^{(t)}, a_i^{(t)}, a_j^{(t-1)} \big| \boldsymbol{m}^{(t-1)}, o^{(t-1)}) \propto P(o^{(t)} \big| a_i^{(t)}) P(a_i^{(t)} \big| a_j^{(t-1)}) P(\boldsymbol{m}^{(t-1)}, o^{(t-1)} \big| a_j^{(t-1)}) P(a_j^{(t-1)}) \quad (5.6)$$

ただし，$P(o^{(t)} \big| a_i^{(t)})$ は動作－物体の関係性を表しており，動作－物体関係モデルより算出する．$P(\boldsymbol{m}^{(t-1)}, o^{(t-1)} \big| a_j^{(t-1)})$ は動作認識の尤度を表し，動作認識モデルを用いて計算する．また，$P(a_i^{(t)} \big| a_j^{(t-1)})$，$P(a_j^{(t-1)})$ はそれぞれ，動作言語モデルの 2-gram と 1-gram である．

5.3.1 動作認識モデル

mMLDA によって分類された時系列データを，それぞれ Multimodal HDP-HMM（MHDP-HMM）[70] で学習し行動モデル集合 $\boldsymbol{A}_M^{\mathrm{K}} \in \{a_M^1, a_M^2, \cdots, a_M^K\}$ を作成する．ただし，行動モデル集合の各要素 a_M^* は行動集合 $\boldsymbol{A}^{\mathrm{K}}$ の a_* と対応している．MHDP-HMM は隠れマルコフモデル（HMM）にディリクレ過程を導入し無限の状態を持つモデルへと拡張した HDP-HMM の各状態から，複数の観測を仮定したモデルである．HDP-HMM の利点としては，状態数を事前に与える必要がない点にある．本章で用いる観測データは，行動する際のユーザの動き（関節角）$\boldsymbol{m}$ と使っている物体 o であり，それぞれガウス分布と多項分布から生成されるものと仮定する．特徴量としては，関節角とその動的特徴を利用する．

学習した行動モデル集合 $\boldsymbol{A}_M^{\mathrm{K}}$ を用いて，入力データ $\boldsymbol{m}_{\mathrm{obs}}$，$o_{\mathrm{obs}}$ に対する行動 a_i の尤度 $P(\boldsymbol{m}_{\mathrm{obs}}, o_{\mathrm{obs}} \big| a_i)$ を計算する．ここで，入力データの時系列は学習データの時系列の途中までであることに注意が必要である．つまり，ユーザの行動が途中であっても次の行動の予測を行う必要がある．

5.3.2 行動言語モデル

行動のパターンをモデル化するために，n-gram 言語モデルを使用する．これは，mMLDA によって記号化された学習データから，各行動の生起回数を数えること

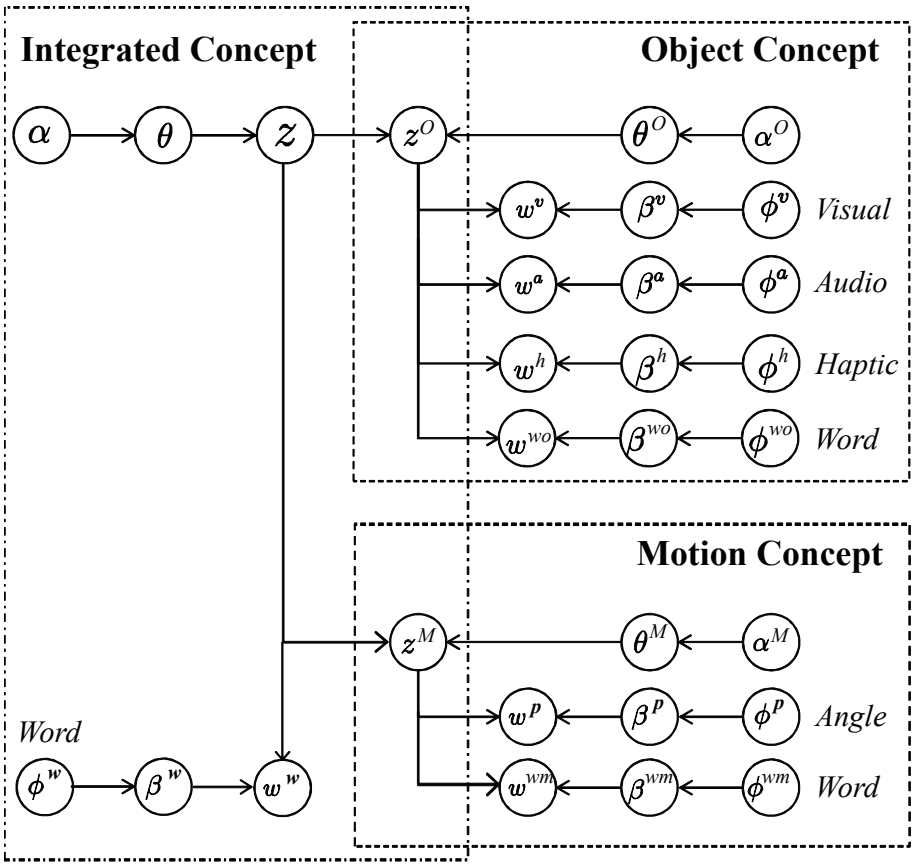

図 5.4: 本章で用いる多層マルチモーダル LDA のグラフィカルモデル

で計算することができる．n をいくつに設定するかは性能や計算量，実際の人の行動がどの程度直前の行動に依存するかによって決まる．本章では，2-gram 及び 1-gram を行動言語モデルとして用いる．

5.3.3 動作ー物体関係モデル

多層 MLDA（mMLDA）は，下位層に物体と動きの分類モデルであるマルチモーダル LDA（MLDA）を，上位層にそれらを統合する MLDA を配置することによって，物体と動きそれぞれのカテゴリ分類を行うと同時に，それらの関連性を教師なしで学習する統計モデルである．図 5.4 に，mMLDA のグラフィカルモデルを示

す．このモデルの学習・認識の詳細については前章を参照されたい．従って，下位層では，何かを飲むような動きといった動作のカテゴリや，ペットボトルという物体のカテゴリが形成され，それらを統合する上位層では，ペットボトルを飲むといった行動が表現される．これは統計モデルであるため，動作の情報を入力することで関連する物体を確率的に予測することができる．

行動 a_i に対する動作－物体の関係性 $P(o|a_i)$ は，学習した mMLDA を用いて以下のように求めることができる．

$$P(o = z^o|a_i = z) = \frac{\alpha^o + N_{z^o z}}{K^o \alpha^o + N_z} \tag{5.7}$$

ただし，z^o と z は mMLDA によって形成された物体概念と上位概念（動作概念）であり，α^o，K^o，$N_{z^o z}$，N_z はそれぞれハイパーパラメータ，カテゴリ数，学習データの全モダリティに対して，物体カテゴリ z^o と上位カテゴリ z が割り当てられた数，上位カテゴリ z が割り当てられた数を表している．

5.4 場所文脈

場所文脈は，ユーザが行動を行う際の位置に対する物体の関係性を表す．ユーザがある場所 $\ell \in \{1, 2, \cdots L\}$ からなる部屋で，ある位置 x において物体 o を使って行動することを想定する場合，場所文脈は以下の式より算出する．

$$P(o|x) = \sum_{\ell} P(x|\ell) P(o|\ell) P(\ell) \tag{5.8}$$

ただし，$P(\ell)$ は一様分布とし，L は部屋内の場所の数であり，$P(x|\ell)$ は場所 ℓ の尤度であり，2 次元ガウス分布で表現する．また，$P(o|\ell)$ は場所 ℓ において，物体 o が使われる確率である．

5.5 音声命令

ユーザは次に必要となるものを，ロボットに音声で命令することが考えられる．その際ロボットは，これを認識し実行する必要がある．しかし一般的な問題として，音声認識は雑音の多い環境では難しい．また，本章では考慮しないが，命令に曖昧性があり解釈を要する場合も少なくない．こうした状況では，上述の文脈情報，つまり現在のユーザの行動と次の行動の予測が役に立つ．

ここでは，物体を届けるタスクのみを想定しており，ユーザは物体名のみを発話すると仮定する．音声命令の具体的な処理は次の通りである．ユーザの音声 S を認識し，上位 D 個の結果 $\boldsymbol{W}^D$ を用いて以下の式を計算する．

$$P(o = z^o \big| S) = \sum_{w^{wo} \in \boldsymbol{W}^D} P(z^o \big| w^{wo}) P(w^{wo} \big| S) \tag{5.9}$$

$$P(z^o \big| w^{wo}) = \frac{1}{P(w^{wo})} \sum_z P(w^{wo} \big| z^o) P(z^o \big| z) P(z) \tag{5.10}$$

$$P(w^{wo}) = \sum_{z^o, z} P(w^{wo} \big| z^o) P(z^o \big| z) P(z) \tag{5.11}$$

$$P(z^o \big| z) = \frac{\alpha^o + N_{z^o z}}{K^o \alpha^o + N_z} \tag{5.12}$$

$$P(w^{wo} \big| z^o) = \frac{\phi^{wo} + N_{z^o wo}}{W^{wo} \phi^{wo} + N_{z^o}} \tag{5.13}$$

ただし，$P(w^{wo} \big| S)$ は音声 S より抽出された物体名 w^{wo} に対する音声認識尤度である．

5.6 実験

ここでは前節で述べた状況を考慮して，提案手法の基礎的な検証を行う．想定したシナリオとして，ロボットが人の行動を観測し，行動するときの人の動き，位置及び動作中に関係する物体情報を KINECT より取得して，ある長さ F フレームのデータを手がかりとして，次の行動に関連する物体を予測する．ただし，ロボットは十分に人の行動を観測し，学習が済んだ段階であるとする．このシナリオを

図 5.5: 実験で使用した物体

表 5.1: 物体に対して行った動き（括弧内はカテゴリ番号）

動き	物体	動き	物体
かける（1）	ドレッシング（3）	拭く（5）	フローリングワイパー（5）
	シャンプー（5）	塗る（6）	スプレー缶（1）
ふる（2）	スプレー缶（1）	見る（7）	ぬいぐるみ（9）
	ペットボトル（2）	置く（8）	カップ麺（4）
	ドレッシング（3）		スナック（7）
飲む（3）	ペットボトル（2）	投げる（9）	ぬいぐるみ（9）
食べる（4）	カップ麺（4）		ガラガラ（10）
	スナック（7）	持ち上げる（10）	クッキー（8）
	クッキー（8）		

実現するために，擬似データを生成した．

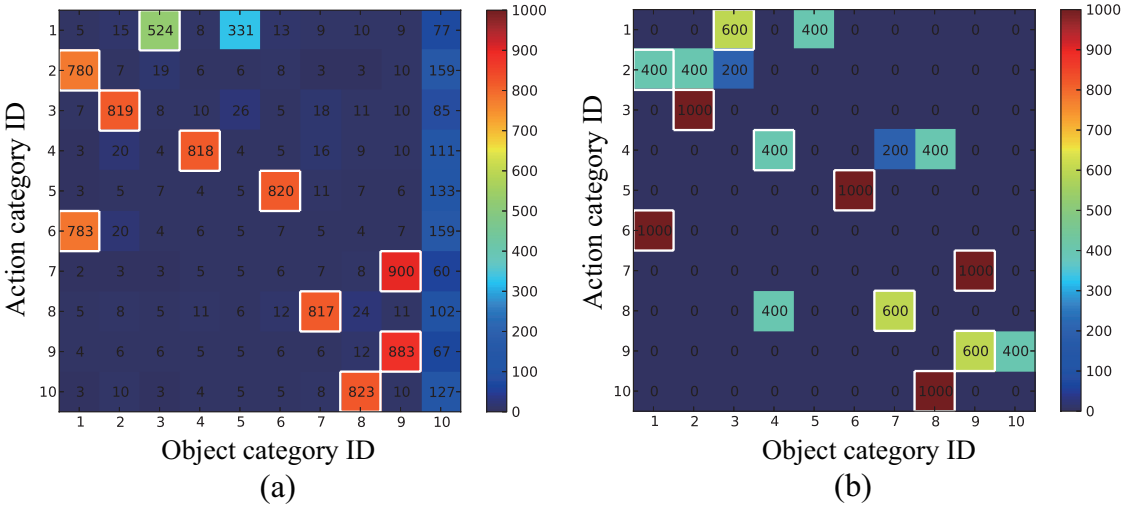

図 5.6: 物体カテゴリと動きカテゴリの共起確率：(a) mMLDA, (b) 正解

5.6.1 擬似データの生成と共起確率

まず，図 5.5 に示すデータセットと表 5.1 の組合せを用いて動作－物体関係モデルと動作認識モデルを作成した．動作－物体関係モデルにおいて，図 5.4 に示す mMLDA を用いた．実際，その結果を図 5.6（a）に示す．これは，$P(o|a)$ の値に相当する．また学習した結果の比較として，表 5.1 に示した各動作の学習サンプルに対して使用される物体の割合を図 5.6（b）にプロットした．図より一見，結果は不一致に見えるが，ここで重要なことは，図 5.6（b）に示した各動作に対して，動作と関係する可能性のある物体を高い確率で予測できることである．これに対して，mMLDA より学習した結果（図 5.6（a）の白い枠）は図 5.6（b）に示した，動作と関係する最も可能性のある物体（図 5.6（b）の白い枠）を高い確率で予測することができるため，その動作と関係する物体を高い確率で当てることができると考えられる．

動作認識モデルにおいて，データセットの各動きに対して，MHDP-HMM のモデルを作成した．ここで，行動集合のサイズ K は表 5.1 の動きの数に合わせて 10 個とした．また，ユーザの部屋内の場所数を $L = 3$ と設定し，各場所 ℓ に対してガウス分布のパラメータを与えた．さらに，図 5.7（a）に示すように各場所と動

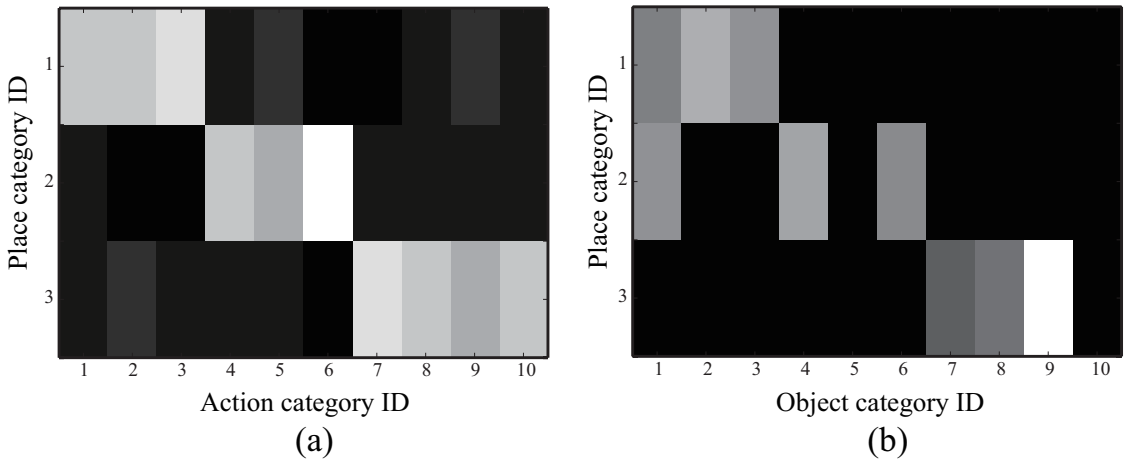

図 5.7: （a）場所カテゴリと行動カテゴリの共起確率,（b）場所カテゴリと物体カテゴリの共起確率

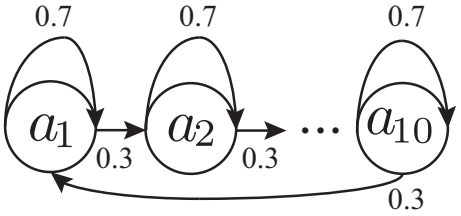

図 5.8: シミュレーション実験に用いた行動の遷移図

作の共起確率を与えた．

次に人の行動の時系列データの生成は，図 5.8 に示すマルコフモデルに従った行動遷移図を用いて行った．図より，各行動 a_* は表 5.1 の動作と対応しており，それぞれ自己遷移確率 0.7 を与えた．ただし，初期遷移は a_1 からとする．このようなパラメータを用いて，2000 個の行動を生成した．図 5.8 より生成された各時系列データに対して，$P(o|a)$ を用いて持ってくる物体を生成する．

音声命令において，表 5.1 のカテゴリ名を物体名とし音声命令を録音した．録音した音声に SNR 100 [dB]，6 [dB]，3 [dB]，0 [dB] の白色雑音をそれぞれ付加した．次に録音した各データに対して，Julius 音声認識エンジンを用いて認識し，上位 5 個の認識尤度を出力した．これらの結果と学習した mMLDA の結果を用いて，音

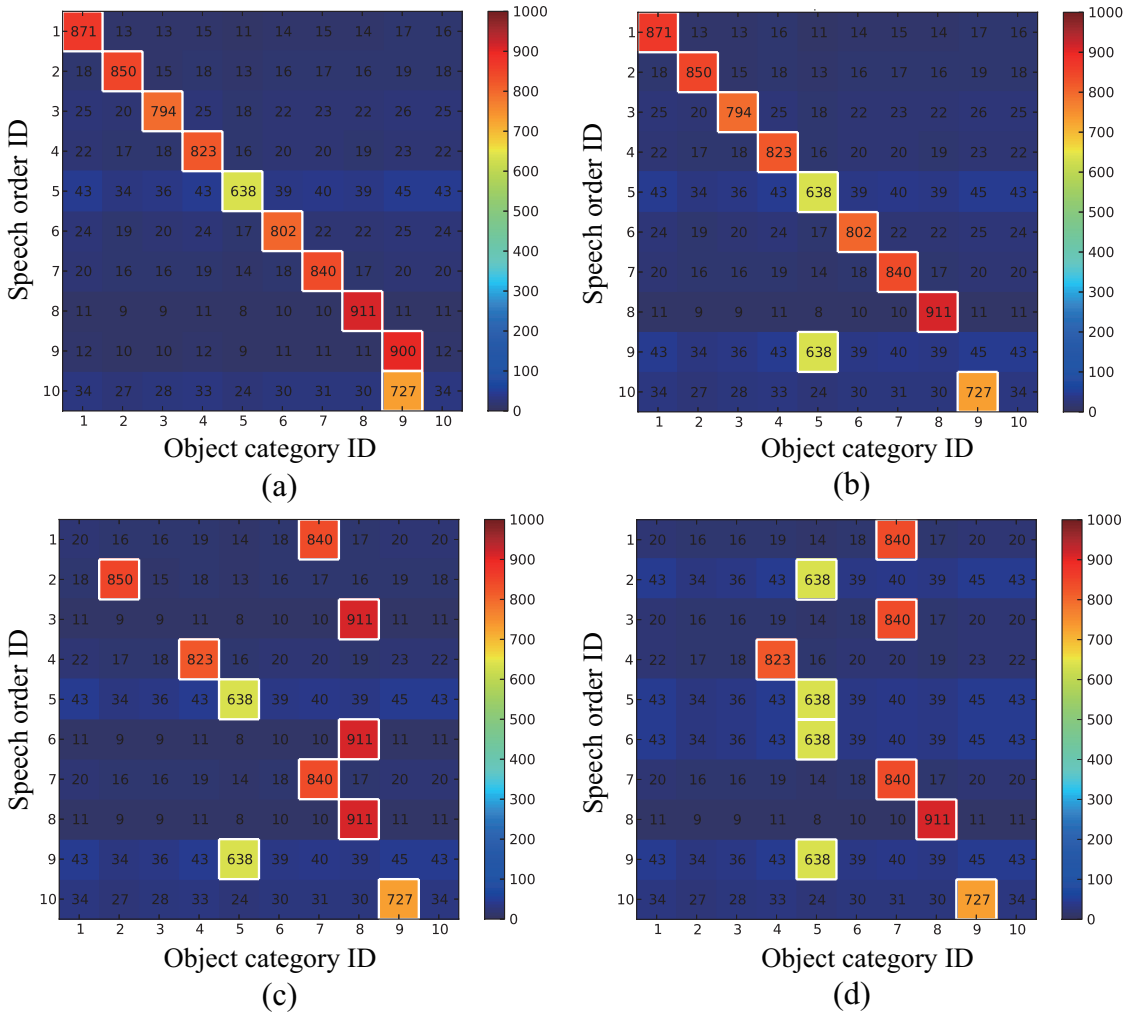

図 5.9: 様々なノイズにおける音声命令と物体カテゴリの共起確率:(a) SNR 100[dB], (b) SNR 6 [dB], (c) SNR 3 [dB], (d) SNR 0 [dB]

声命令と物体の関係性を求めた結果を図 5.9 に示す．図より，クリーンな環境で認識した場合，物体カテゴリ 10 以外は全て正しく想起することができたが，SNR が低くなるに連れて結果が悪くなっていることが分かる．ノイズがない場合に対する誤りの原因として，mMLDA によって分類された物体カテゴリ 10（ガラガラ）が物体カテゴリ 9（ぬいぐるみ）に分類されてしまったことが考えられる．

最後に，生成された時系列データを学習用と認識用に分割した．学習用のデー

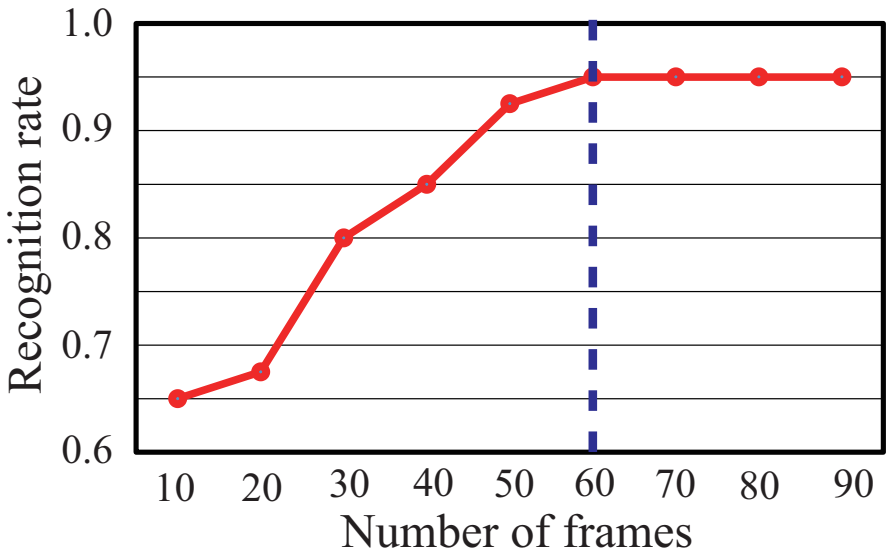

図 5.10: 観測されたフレーム数に対する動作認識率

タを用いて，動作言語モデルと場所に対する物体頻度（図 5.7（b））をそれぞれ計算した．また，行動文脈，場所文脈，音声文脈を計算し，それらを一つのヒストグラムとして SVM で学習した．学習したパラメータを用いて，認識用のデータを用いて行動文脈，場所文脈，音声文脈を計算し，学習と同様な方法を用いて認識した．

5.6.2 実験結果

まず動作認識におけるフレームの長さ F の影響を検討するために，動作認識モデルを用いて認識用のデータで動作認識を行った．図 5.10 より，F を 60 に設定すれば，95%の認識率が得られることが分かる．以降ロボットの行動決定実験には，この値を用いる．その結果を図 5.11 に示す．図より，全ての SNR に対して平均した結果について，単一の文脈を用いる場合は 70%以下の認識率となることが分かる．一方，SVM を用いて文脈を統合した場合，94.2%まで認識率が向上した．従って，行動文脈や場所文脈など様々な文脈を統合することでよりロバストな行動決定が行えると言える．

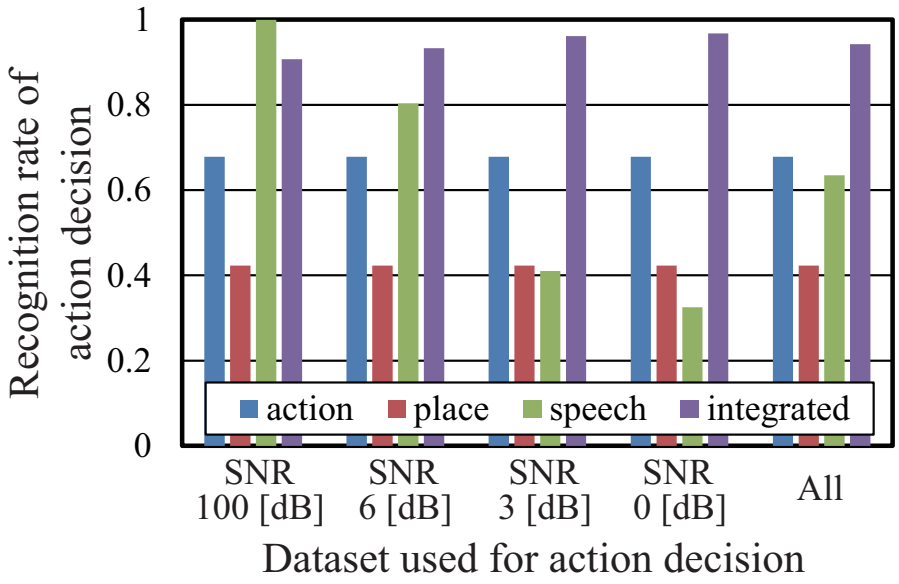

図 5.11: ロボットの行動決定結果

5.7 まとめ

ロボットは人を相手にして作業あるいはサービスを行うために，背景知識や周辺の状況などといった文脈を考慮しなければ適切に行うことができない．本章において，ロボットが人と生活する上で，様々な文脈の中でどのように行動すれば良いかを議論した．特に人の活動に対して人の習慣（現在の動作と次の動作と動作する際に関係する物体や場所など）が最も関係している．前章までで提案した mMLDA によって，人の動作に関係する物体や動きなどの共起性を手がかりとして動作概念の獲得を行うことができることを示した．しかし，人の習慣において現在の動作と次における動作，いわば推移性を手がかりとしていないため，このようなことを表現することができなかった．

本章では，mMLDA の応用として，推移性を手がかりとした行動文脈を提案し，場所文脈及び音声命令と統合した，ロボットの行動決定手法を提案した．提案手法ではロボットがユーザの生活に密着し行動を観測することで，行動パターンを学習する．学習したユーザの行動パターンを利用して，次の行動を予測し，予測

した次の行動を文脈として音声命令，場所情報と統合することで最終的なサービス行動を決定する．本章では，全ての文脈を確率的枠組みで定義し，SVM を用いて統合する手法を提案した．提案アルゴリズムの基礎的な検討として，ノイズのある環境でロボットの行動決定の実験を行い，その実現可能性を示した．

本章において，人の動きと物体の関係を mMLDA で表現し行動文脈として用いるが，場所文脈とは別モデルとして扱っている．しかし前章で述べたように，場所と動作の関係を全て mMLDA で表現することが可能である．さらに，場所と動作だけではなく人物などの関係を mMLDA でモデル化することで，それらを予測することが可能となり，ロボットが行えるサービスも物体に対するものだけではなく，場所の案内なども考えられる．このように，多様な概念を mMLDA でモデル化しそれと音声命令などを統合することが今後の課題である．

さらに，本章では提案手法において基礎的な検証を行ったが，より現実的なシナリオやフィールドでの学習・行動決定実験も今後の課題として挙げられる．著者は，例えばお茶の水女子大学の実験住宅である OCHA ハウス [72] での実験を考えている．ここでの実験において，ロボットによる能動的なセンシングを考慮し，実際にロボットによるサービスを実現したい．

第6章　まとめ

6.1　まとめ

本論文では，ロボットが自身の身体性を利用して得た経験を通して多様な概念を獲得することができるモデルを提案した．提案モデルはマルチモーダルカテゴリ分類が可能な確率モデルである MLDA を多層化する mMLDA である．mMLDA の下位層は，物体，動き，場所と人物概念がそれぞれ表現され，それらの関係を表す上位概念は上位層に表される．形成された多様な概念を用いることで，未観測情報に対する予測や概念間における予測などを確率的に行うことが可能となった．また，階層的な概念における単語の結び付け方法も提案した．提案手法では，相互情報量を用いて単語の各概念に対する重みを計算しそれを単語予測に利用することで，相互情報を用いない単語予測結果に比べ精度が向上することを実験を通して示した．さらに，教示文に含まれる単語列に対して，各単語が指し示す概念の遷移を求めることで文法の学習を可能とする手法を提案した．こうした多様な概念，それらを指している単語と獲得した文法を用いることで，観測情報から文章の生成を行うことが可能となった．

本論文ではさらに，人の活動における行動と行動の時間的な関係を表現することが可能な手法を提案した．提案手法では，各行動は mMLDA によって表現され，行動と行動の遷移を計算し，これらを行動文脈とする．行動文脈を考慮することで，次の行動の予測だけではなく，その行動に関係する物体の予測が可能となった．また行動文脈を音声命令など様々な文脈と統合したロボットによる行動決定手法も提案し，基礎的な検証を通して可能性を示した．

MLDA における学習は，ロボットがセンシングできる視覚，聴覚や触覚などといった感覚情報の共起を用いて実現していると考えることができる．これに対し

て提案した mMLDA は，感覚情報の共起だけではなく，それらを抽象化する物体，動きや場所などの概念の共起性を手がかりとして学習する．これらの情報は実際に，我々が日々行っている活動に含まれていると考える．さらに人の活動においては，行動と行動の順序といった時間的な情報（推移性）も含んでおり，ロボットにおいて真の理解を実現するための欠かせない情報である．本論文では，共起性と推移性を手がかりとした確率的な知識を表現し，それをロボットが自律的に獲得することのできる枠組みを提案した．ロボットが人のように知的に振る舞うためには知識が必要であり，自身の身体性を活かして知識を獲得することこそが，真の理解につながるのではないかと考えている．

6.2 掃除タスクはどこまで可能か

掃除タスクは 2 章で取り上げた家庭環境におけるタスクの一例であり，一般的に汚い状態をきれいな状態にするタスクであると考えられる．ここでまず,「きれい」や「汚い」という概念の判断ができるとすると，どのようにそれらの状態を変化させるのかという行動決定の問題となる．行動の決定を行うとき，現在の状況を把握する必要があり，それに対する行動を選択しなければならない．ここでは，理想的な掃除タスクを基準にして，タスクを行うためのロボットの能力として要求されるものを表 6.1 にまとめ，2 章で述べた掃除タスク（以下「作り込み」と呼ぶ）と提案した mMLDA を用いた掃除タスク（以下「mMLDA」と呼ぶ）の実現可能性を比較する．

表 6.1 より，理想的な掃除タスクを行うためにはまず，掃除するための道具に対して予測する能力が必要とされる．これは，見たことのない道具に対して，どんな道具なのかどのようにすればその道具が使えるのかなどといった予測をする必要があるためである．例えば，新しく発売され形が独特な「掃除機」を見た場合，その物体が掃除機であり，細かいごみの上で動かせばごみを吸うことができるといった推論を行える．この能力は,「作り込み」の場合は当然発揮できないが,「mMLDA」では獲得された知識の範囲であれば実現することが可能である．実際 mMLDA を用いることで,「掃除機」という物体概念と「何かの上で動かす」とい

表 6.1: 掃除タスクのためのロボットの能力の実現可能性の比較

No	ロボットの能力	理想	作り込み	mMLDA
1	道具に対する予測	○ 可能	× 不可能	△ 獲得された知識の範囲において可能
2	人の命令に対する理解	○ 可能	△ キーワードマッチングより簡易的に可能	△ 獲得された言語であれば可能
3	未知環境に対する行動	○ 可能	× 不可能	△ 獲得された知識の範囲において可能
4	教師なしで学習	○ 可能	× 不可能	△ 可能
5	掃除タスクに必要な概念を全て持っている	○ 持っている	× 持っていない	× 全ては持っていない
6	行動生成とプランニング	○ 可能	△ 想定環境であれば可能	× できていない

う動き概念が関係しており,「掃除機をかける」という行動概念が形成できることは明らかである．これらの概念がロボットの知識となり，この知識を利用することで様々な予測が可能となる．

次に，どのタスクにおいても人の命令を理解する能力が必要である．ここでの理解の定義は，前述のように概念を通して様々な予測ができることである．この定義に従えば,「作り込み」の場合キーワードマッチングを行うことで簡易的に人の命令を理解することができる．しかしこれは，命令を単なる記号として扱うに過ぎない．これに対して「mMLDA」を用いることで，言語の獲得を行うことが可能となるため，獲得された言語であれば本質的な理解を行うことが可能となる．実際,「掃除しろ」という命令に対して「mMLDA」を用いると,「掃除」という単語がどの概念と結び付いているのかが分かる．それらの概念と対応する行動を実行すれば，タスクを行うことができると考えられる．

また，未知の環境に対して行動することができることも柔軟なタスクを行える上では欠かせない能力の一つであろう．これに対して,「作り込み」では想定された環境でしか動作することができないため，行動することは不可能である．一方「mMLDA」では，獲得された知識の範囲であれば，上の例で述べたように未知の環境に対して様々な推論をすることが可能となり，それをもとに行動することもできる．さらに,「mMLDA」では教師なしで学習することができるため,「作り込み」のように人が予め与えていなくても，自身の経験を通して自律的に学習することもできる．

しかし表 6.1 に示すように,「mMLDA」は掃除タスクに必要な概念を全てを持っ

ている訳ではない．しかしこれはモデルを拡張することで解決できる可能性がある．一方,「作り込み」の場合，そもそも掃除に対する概念を持っていないため，対処することが困難であると考えられる．また，2 章で述べたように「作り込み」において，行動生成とプランニングは想定された環境であれば行うことができる．これに対して，現状の「mMLDA」では実現できていない．さらに，掃除タスクにおける「きれい」や「汚い」という概念，いわば感性の仕組みもまだ実現できておらず，今後検討していく必要がある．

6.3 今後の課題

6.3.1 タスクに対する知識の利用

以上の議論を踏まえてここではロボットが獲得した知識を用いて，どのようにタスクに活用できるかについて考察したい．ロボットのタスクは，ロボット自身が置かれた環境との相互作用を通して遂行されると考えられる．当然のことながら，タスクをこなすためには適切な行動を選択しなければならず，これは実際にタスクを実行するために行動決定問題を考える必要があることを意味している．行動決定問題を自律的にロボットが実現するための枠組みとして，強化学習を用いるのが一般的である [73]．強化学習では，ロボットがある状態において行動を行うことで環境が変化し，それに対応した報酬が与えられる．この報酬は，ある状況においてロボットがとった行動に対する評価となる．この枠組では，報酬の最大化としてタスクにおける行動決定問題を解くことができる．この際重要なのは，状態空間の決定，行動の設定，状態の認識，報酬の設定である．本論文で扱ったのは，状態空間の決定，行動の設定，状態の認識の問題である．図 2.20 に示したように，知覚情報がカテゴリ分類されることで状態空間が生成され，さらにはこれらと動きの情報との関係から行動セットも行動概念として自動的に獲得されることになる．さらには，知覚情報からこれらの状態や行動を推論することができるため，状態の認識も可能な枠組みとなっている．しかしこの枠組みには報酬が含まれていないため，これだけでは強化学習を実現することはできない．一般に報酬は，タスクによって異なり，基本的には設計者が設定するのが一般的である．

著者は報酬の設定に関して，例えば人とのインタラクションやロボットに感情 [74] を与えることで実現できると考えているが，これについては今後の課題であると考えている．

また本来人間は感情や感性を持っており，感性によって知識や行動が影響を受ける．ここでは，ロボットの感性は知覚情報のフィルタとしての役割を果たすと考える．文献 [75] では，ロボットの感性として「美しい」をどのように考えることができるかが議論されている.「美しい」とは，ある概念の中心的なものに感じる感覚であると考えられ，非常に予測の精度が高いものに感じると考えるのが妥当であろう．この仕組みは，入力された知覚情報を概念構造に当てはめたときに，その情報がある概念の中心に来るような場合のメタ認知として解釈することができる．著者はこのような仕組みをロボットに持たせることができれば，掃除タスクにあったような「きれい」や「汚い」という概念を実現することができるのではないかと考えており，このような仕組みを今後検討する必要があると考えている．

6.3.2 提案モデルに対する課題

提案モデルでは，確率モデルであるマルチモーダル LDA（MLDA）を階層化した多層マルチモーダル LDA（mMLDA）であり，各概念に対してカテゴリ数を予め人手で与えなければならない．しかし，実際にロボットが階層的な概念を形成する際に，予めカテゴリ数を得ることはできない．また，環境やセンサの性能によって取得されるマルチモーダル情報が異なるため，適切なカテゴリ数を事前に与えることは困難である．従って，各概念におけるカテゴリ数を自動的に推定可能なモデルを考える必要があると言える．この問題に対する解決方法として，変分ベイズ法やノンパラメトリックベイズなどが挙げられる．例えば，先行研究ではノンパラメトリックベイズである Hierarchical Dirichlet Processes（HDP）[60] をマルチモーダルに拡張したマルチモーダル HDP（MHDP）[62] が提案されており，物体カテゴリ分類に対してその有効性を示した．この知見を活かして，MHDP を階層化することでカテゴリ数を自動的に推定可能なモデルを今後検討する必要があると考える．

また本論文における言語生成は，獲得した文法と言語モデルを統合する手法に

基づいている．この手法では学習したシーンに対する文を生成することはできるが，未学習のシーンに対しての生成が困難である．これを防ぐために，今後の課題として概念クラスのバイグラムと統語を用いて，文法の学習と文生成を行うことを考えている．

また本論文では，実際のコミュニケーションに重要となる様々な文脈をモデル化し，それをロボットに応用する手法を提案した．提案手法では行動文脈，場所文脈及び音声命令を統合してロボットの行動決定を実現し，シミュレーション実験でその有効性を示した．今後は，より現実的なシナリオやフィールドでの学習・行動決定実験を行う予定である．

参考文献

[1] iROBOT，“ROOMBA”，http://www.irobot.com/.

[2] SONY，“AIBO”，http://www.sony.jp/products/Consumer/aibo/.

[3] NDSOFT，“PARO”，http://www.ndsoft.jp/paro.php/.

[4] HONDA，“ASIMO”，http://www.honda.co.jp/ASIMO/.

[5] TOYOTA，“TPR”，http://www.toyota.co.jp/.

[6] DARPA，“DARPA Robotics Challenge”，http://www.theroboticschallenge.org/.

[7] “Robocup@home”，http://www.ai.rug.nl/robocupathome/.

[8] J. Locke，“An Essay Concerning Human Understanding”，London，1689.

[9] D. L. Medin，and L. J. Rips，“Concepts and Categories: Memory，Meaning，and Metaphysics”，2005.

[10] F. G. Ashby，and W. T. Maddox，“Human Category Learning”，Annual Review of Psychology，vol. 56，pp. 149–178，2005.

[11] E. Rosch，“Principles of categorization”，Concepts: core readings，pp. 189–206，1999.

[12] S. Lewandowsky，M. Kalish，S. K. Ngang，“Simplified learning in complex situations: knowledge partitioning in function learning”，Journal Exp. Psychol: Gen，vol. 131，pp. 163–193，2002.

[13] A. B. Markman，B. H. Ross，“Category use and category learning”，Psychol. Bull, vol. 129，no. 4，pp. 592–613，2003.

[14] E. M. Markman，“The whole-object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings”，Perspectives on Language and Thought: Interrelations in Development，pp. 72–106，1991.

[15] S. Harnad，“The symbol grounding problem”，Physica D，vol. 42，pp. 335–346，1990.

[16] J. F. Sowa，“Semantic Networks”，Encyclopedia of Artificial Intelligence，Wiley，1987.

[17] S. J. Russell，P. Norvig，“Artificial intelligence : a modern approach (3rd ed.)”. Prentice Hall，2010.

[18] 中村友昭，長井隆行，岩橋直人，“ロボットによる物体のマルチモーダルカテゴリゼーション”，電子情報通信学会和文論文誌，vol. J92-D，no. 10，pp. 2507–2518，2008.

[19] R. Fergus，P. Perona，and A. Zisserman，“Object Class Recognition by Unsupervised Scale-Invariant Learning”，in Proc. of CVPR 2003，vol. 2，pp. 264–271，2003.

[20] J. Sivic，B. C. Russell，A. A. Efros，A. Zisserman，and W. T. Freeman，“Discovering Object Categories in Image Collections”，in Proc. of ICCV 2005，pp. 370–377，2005.

[21] L. Fei-Fei，“A bayesian hierarchical model for learning natural scene categories”，IEEE Conf. on Computer Vision and Pattern Recognition，pp. 524–531，2005.

[22] C. Wang，D. Blei. and L. Fei-Fei，“Simultaneous image classification and annotation”，IEEE Conf. on Computer Vision and Pattern Recognition，pp. 1903–1910，2009.

[23] E. Torres-Jara, L. Natale, and P. Fitzpatrick, “Tapping into Touch”, Lund University Cognitive Studies, pp. 22–24, 2005.

[24] J. Sinapov, and A. Stoytchev, “Object Category Recognition by a Humanoid Robot Using Behavior-grounded Relational Learning”, in Proc. of ICRA 2011, pp. 184–190, 2011.

[25] W. Takano, H. Imagawa, and Y. Nakamura, “Prediction of Human Behaviors in the Future through Symbolic Inference”, in Proc. of ICRA 2011, pp. 1970–1975, 2011.

[26] W. Takano, and Y. Nakamura, “Bigram-Based Natural Language Model and Statistical Motion Symbol Model for Scalable Language of Humanoid Robots”, in Proc. of ICRA 2012, pp. 1232–1237, 2012.

[27] T. Taniguchi, and S. Nagasaka, “Double Articulation Analyzer for Unsegmented Human Motion using Pitman-Yor Language Model and Infinite Hidden Markov Model”, in Proc. of SII 2011, pp. 250–255, 2011.

[28] T. Ogata, S. Nishide, H. Kozima, K. Komatani, and H. Okuno, “Intermodality Mapping in Robot with Recurrent Neural Network”, Pattern Recognition Letters, vol. 31, no. 12, pp. 1560–1569, 2010.

[29] L. Montesano, M. Lopes, A. Bernardino, and J. S. Victor, “Learning Object Affordances: From Sensory-Motor Coordination to Imitation”, IEEE Trans. on Robotics, vol. 24, no. 1, 2008.

[30] B. Moldovan, P. Moreno, M. Otterlo, J. S. Victor, and L. D. Raedt, “Learning Relational Affordance Models for Robots in Multi-Object Manipulation Tasks”, in Proc. of ICRA 2012, pp. 4373–4378, 2012.

[31] A. Gupta, A. Kembhavi, and L. S. Davis, “Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition”, IEEE Trans. on PAMI, vol. 31, no. 10, pp. 1775–1789, 2009.

[32] B. Yao, and L. Fei-Fei, "Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses", IEEE Trans. on PAMI, vol. 34, pp. 1691–1703, 2012.

[33] C. L. Teo, Y. Yang, H. Daumé III, C. Fermüller, Y. Aloimonos, "A Corpus-Guided Framework for Robotic Visual Perception", in Proc. of AAAI 2011, 2011.

[34] H. Yu, J. M. Siskind, "Grounded Language Learning from Video Described with Sentences", in Proc. of ACL, pp. 53–63, 2013.

[35] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, "Grounding Action Descriptions in Videos", in Proc. of ACL, pp. 25–36, 2013.

[36] R. Brooks, "A robust layered control system for a mobile robot", IEEE Journal of Robotics and Automation, vol. RA-2, no. 1, 1986.

[37] M. Attamimi, A. Mizutani, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, "Real-Time 3D Visual Sensor for Robust Object Recognition", Int. Conf. on IROS, pp. 4560–4565, 2010.

[38] S. M. Lavalle, "Rapidly-exploring random trees: A new tool for path planning", 1998.

[39] H. Murase, V. V. Vinod, "Fast Visual Search Using Focussed Color Matching: Active Search", IEIEC J81-D-2, no. 9, pp. 2035–2042, 1998 (in Japanese).

[40] K. Okada, S. Kagami, M. Inaba, H. Inoue, "Plane Segment Finder: Algorithm, Implementation and Applications", Int. Conf. on ICRA, pp. 2120–2125, 2001.

[41] R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin, "Shape Distributions", ACM Transactions on Graphics, vol. 21, no. 4, pp. 807–832, 2002.

[42] G. Csurka, C. Dance, L. Fan, J. Williamowski, C. Bray, “Visual Categorization with Bags of Keypoints”, Int. Workshop on Statistical Learning in Computer Vision, pp. 1–22, 2004.

[43] E. Nowak, F. Jurie, B. Triggs, “Sampling Strategies for Bag-of-Features Image Classification”, Int. Conf. on ECCV, vol. 3954, pp. 490–503, 2006.

[44] A. Vedaldi, B. Fulkerson, “VLFeat-An Open and Portable Library of Computer Vision Algorithms”, ACM Multimedia, pp. 1469–1472, 2010.

[45] T. Oggier, “Miniature 3D TOF Camera for Real-Time Imaging”, Perception and Interactive Technology, pp. 212–216, 2006.

[46] M. Bohme, “Shading Constraint Improves Accuracy of Time-of-Flight Measurements”, CVIU, 2010.

[47] M. Sturmer, “Standardization of Intensity Values Acquired by Time-of-Flight-Cameras”, CVPRW, 2008.

[48] C. C. Chang, and C. J. Lin, “LIBSVM: A Library for Support Vector Machines”, ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pp. 1–27, 2011.

[49] C. E Rasmussen, “The Infinite Gaussian Mixture Model”, In Advances in Neural Information Processing Systems, vol. 12, pp. 554–560, 2000.

[50] D. G*örü*r, and C. E. Rasmussen, “Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution”, Journal of Computer Science and Technology, vol. 25, no. 4, pp. 653–664, 2010.

[51] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, “The Pascal Visual Object Classes (VOC) Challenge”, Int. Journal of Computer Vision, vol. 88, no. 2, pp. 303–338, 2010.

[52] P. J. Besl, and N. D. McKay, “A Method for Registration of 3-D Shapes”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 14, no. 2, pp. 239–256, 1992.

[53] M. Attamimi, T. Nakamura, and T. Nagai, “Hierarchical Multilevel Object Recognition Using Markov Model”, in. Proc. ICPR 2012, pp. 2963–2966, 2012.

[54] M. Attamimi, K. Ito, T. Nakamura, and T. Nagai, “A Planning Method for Efficient Mobile Manipulation Considering Ambiguity”, in. Proc. IROS 2012, pp. 965–972, 2012.

[55] 長井隆行, 中村友昭, “マルチモーダルカテゴリゼーション : 経験を通して概念を形成し言葉の意味を理解するロボットの実現に向けて”, 人工知能学会誌, vol. 27, no. 6, pp. 555–562, 2012.

[56] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis”, Machine Learning, vol. 42, pp. 177–196, 2001.

[57] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation”, Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[58] T. Nakamura, T. Araki, T. Nagai, and N. Iwahashi, “Grounding of Word Meanings in LDA-Based Multimodal Concepts”, Advanced Robotics, vol. 25, no. 17, pp. 2189–2206, 2011.

[59] 中村友昭, 西田匡志, 長井隆行, “把持動作による物体カテゴリの形成と認識”, 情報処理学会全国大会, 5V-3, 2010.

[60] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical Dirichlet processes”, Journal of the American Statistical Association, vol. 101, no. 476, pp. 1566–1581, 2006.

[61] O. Mangin and P. Y. Oudeyer, “Learning to Recognize Parallel Combinations of Human Motion Primitives with Linguistic Descriptions using Non-negative Matrix Factorization”, in Proc. of IROS 2012, pp. 3268–3275, 2012.

[62] 中村友昭，荒木孝弥，長井隆行，岩橋直人，“階層ディリクレ過程に基づくロボットによる物体のマルチモーダルカテゴリゼーション”，計測自動制御学会論文集，pp. 469–478，vol. 49，no. 4，2013.

[63] K. Kinoshita, Y. Konishi, S. Lao, and M. Kawade, “Facial Feature Extraction and Head Pose Estimation Using Fast 3D Model Fitting”, in Proc. of MIRU 2008, pp.1325–1329, 2008 (in Japanese)

[64] Y. Konishi, K. Kinoshita, S. Lao, and M. Kawade, “Real-Time Estimation of Smile Intensities”, in Proc. of Interaction 2008, no. 2008, vol. 4, pp.47–48, 2008 (in Japanese)

[65] K. Papineni, S. Roukos, T. Ward, W. J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation”，in Proc. of ACL，2002.

[66] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, “The infinite hidden markov model”，Advances in neural information processing systems，pp. 577–584，2001.

[67] R. Kelley, M. Nicolescu, A. Tavakkoli, C. King, and G. Bebis, “Understanding human intentions via Hidden Markov Models in autonomous mobile robots”，ACM/IEEE Int. Conf. in HRI，pp. 367–374，2008.

[68] D. Gehrig, P. Krauthausen, L. Rybok, H. Kuehne, U. D. Hanebeck, T. Schultz, and R. Stiefelhagen, “Combined intention, activity, and motion recognition for a humanoid household robot”，IEEE Int. Conf. on IROS，pp. 4819–4825，2011.

[69] H. Koppula, R. Gupta, and A. Saxena, “Learning Human Activities and Object Affordances from RGB-D Videos”, Int. Journal of Robotics Research, vol. 32, no. 8, pp. 951–970, 2013.

[70] 中村友昭，船越孝太郎，長井隆行，“HDP-HMM を用いたロボットによる物体軌道の学習と予測”，日本ロボット学会学術講演会，2C1-05，2013.

[71] K. Sugiura, N. Iwahashi, H. Kawai, and S. Nakamura, “Situated spoken dialogue with robots using active learning”, Advanced Robotics, vol. 25, no. 17, pp. 2207–2232, 2011.

[72] 太田裕治，元岡展久，椎尾一郎，塚田浩二，神原啓介，“ユビキタスコンピューティング実験住宅における無侵襲歩行モニタリングの試み”，電気学会論文誌 C 編（電子・情報・システム部門誌）vol. 130, no. 3, pp. 383–387, 2010.

[73] 浅田稔，“ロボットの行動獲得のための能動学習”，情報処理，vol. 38, no. 7, pp. 583–588, 1997.

[74] 山口拓郎，アッタミミ ムハンマド，中村友昭，長井隆行，池原雅章，“マルチモーダル LDA を用いたロボットによる感情語彙の獲得”，第 14 回計測自動制御学会システムインテグレーション部門講演会，1M4-5，2014.

[75] 長井隆行，“自身の経験が生み出すロボットの知性と感性”，第 15 回感性工学会大会, 企画：感性ロボティクスの未来 (感性ロボティクス部会)，招待講演，D71，2013.

発表実績

関連論文

論文

[1] Muhammad Attamimi，Takaya Araki，Tomoaki Nakamura，and Takayuki Nagai，“Visual Recognition System for Cleaning Tasks by Humanoid Robots”，International Journal of Advanced Robotic Systems: Humanoid, pp.1–14, 2013.（**2 章の内容に関連**）

[2] アッタミミムハンマド，ファドリルムハンマド，阿部香澄，中村友昭，船越孝太郎，長井隆行，“多層マルチモーダル LDA を用いた人の動きと物体の統合概念の形成”，日本ロボット学会誌，vol.32, no.8，pp.753–764，2014.（**3 章の内容に関連**）

国際会議プロシーディングス

[1] Muhammad Attamimi, Tomoaki Nakamura, and Takayuki Nagai, “Hierarchical Multilevel Object Recognition Using Markov Model”, ICPR 2012, pp.2963–2966, 2012.（**2 章の内容に関連**）

[2] Muhammad Attamimi, Muhammad Fadlil, Kasumi Abe, Tomoaki Nakamura, Kotaro Funakoshi, and Takayuki Nagai, “Integration of Various Concepts and Grounding of Word Meanings Using Multi-layered Multimodal LDA for Sentence Generation”, IROS 2014，pp.2194–2201，2014.（**4 章，5 章の内容に関連**）

参考論文

論文

[1] 中村友昭，アッタミミムハンマド，杉浦孔明，長井隆行，岩橋直人，岡田浩之，大森隆司，“拡張モバイルマニピュレーションのための新規物体の学習”，日本ロボット学会誌，vol.30，no.2，pp.213–224，2012.

[2] Miriam Lopez-de-la-Calleja, Takayuki Nagai, Muhammad Attamimi, Mariko Nakano-Miyatake, and Hector Perez-Meana, “Object Detection Using SURF and Superpixels”, Journal of Software Engineering and Applications, vol.6, no.9，pp.511–518, 2013.

[3] 阿部香澄，日永田智絵，アッタミミ ムハンマド，長井隆行，岩崎安希子，下斗米貴之，大森隆司，岡夏樹，“人見知りの子どもとロボットの良好な関係構築に向けた遊び行動の分析”，情報処理学会論文誌，vol.55，no.12，pp.2524–2536，2014.

国際会議プロシーディングス

[1] Muhammad Attamimi, Akira Mizutani, Tomoaki Nakamura, Komei Sugiura, Takayuki Nagai, Naoto Iwahashi, Hiroyuki Okada, and Takashi Omori，“Learning Novel Objects Using Out-of-Vocabulary Word Segmentation and Object Extraction for Home Assistant Robots”, ICRA 2010, pp.745–750, 2010.

[2] Muhammad Attamimi, Akira Mizutani, Tomoaki Nakamura, Takayuki Nagai, Kotaro Funakoshi, and Mikio Nakano, “Real-Time 3D Visual Sensor for Robust Object Recognition”, IROS 2010, pp.5410–5415, 2010.

[3] Mikio Nakano, Naoto Iwahashi, Takayuki Nagai, Taisuke Sumii, Xiang Zuo, Ryo Taguchi, Takashi Nose, Akira Mizutani, Tomoaki Nakamura,

Muhammad Attamimi, Hiromi Narimatsu, Kotaro Funakoshi, Yuji Hasegawa, “Grounding New Words on the Physical World in Multi-Domain Human-Robot Dialogues”, AAAI 2010, pp.74–79, 2010.

[4] Muhammad Attamimi, Keisuke Ito, Tomoaki Nakamura, and Takayuki Nagai, “A Planning Method for Efficient Mobile Manipulation Considering Ambiguity”, IROS 2012, pp.965–972, 2012.

[5] Muhammad Attamimi, Kasumi Abe, Akiko Iwasaki, Takayuki Nagai, Takayuki Shimotomai, and Takashi Omori, “Robots That Can Play with Children: What Makes a Robot Be a Friend”, ICONIP 2013, vol.8226, pp.377–386, 2013.

[6] Chie Hieida, Kasumi Abe, Muhammad Attamimi, Takayuki Shimotomai, Takayuki Nagai, and Takashi Omori, “Physical Embodied Communication between Robots and Children: An Approach for Relationship Building by Holding Hands”, IROS 2014, pp.3291–3298，2014.

[7] Kasumi Abe, Chie Hieida, Muhammad Attamimi, Takayuki Nagai, Takayuki Shimotomai, Takashi Omori, and Natsuki Oka, “Toward Playmate Robots that can Play with Children Considering Personality”, HAI 2014, pp.165–168, 2014.

口頭発表

[1] 水谷了，Attamimi Muhammad，中村友昭，長井隆行，船越孝太郎，中野幹生, “赤外線 TOF カメラと CCD カメラのキャリブレーションによる 3 次元センサの実現と画像処理への応用”，電気学会計測研究会，pp.69–74，2009.

[2] Attamimi Muhammad，水谷了，中村友昭，長井隆行，船越孝太郎，中野幹生, “複数特徴量を統合したパーティクルフィルタによる物体の検出と追跡,” 情報処理学会全国大会，5Y-1，2010.

[3] 水谷了，中村友昭，Attamimi Muhammad，長井隆行，船越孝太郎，中野幹生, “距離情報を用いた3次元物体認識,” 情報処理学会全国大会，3X-9，2010.

[4] Attamimi Muhammad，丸山恭平，前田泰斗，中村友昭，長井隆行，“物体認識と材質認識を用いた掃除タスクの実現”，日本ロボット学会学術講演会，2J1-2，2011.

[5] Attamimi Muhammad，中村友昭，長井隆行，“ロボットによる掃除タスクのための視覚認識システム”，創発システム・シンポジウム ポスター発表，pp.67–70，2011.

[6] Attamimi Muhammad，中村友昭，長井隆行，“近赤外線反射強度を用いた材質の認識とその応用”，信学技報，vol.111，no.257，SIP2011-69，pp.43–48，2011.

[7] Attamimi Muhammad，中村友昭，長井隆行，“マルコフモデルに基づく階層物体認識”，創発システム・シンポジウム ポスター発表，p.20，2012.

[8] Attamimi Muhammad，中村友昭，長井隆行，“マルコフモデルを用いた階層物体認識”，日本ロボット学会学術講演会，2J1-5，2012.

[9] Attamimi Muhammad，中村友昭，長井隆行，“サービスロボットの遠隔操作からの自律化”，日本ロボット学会学術講演会，2C2-04，2013.

[10] Muhammad Fadlil，Attamimi Muhammad，長井隆行，中村友昭，船越孝太郎，“多層マルチモーダルLDAと相互情報量による語意の獲得”，日本ロボット学会学術講演会，2C2-06，2013.

[11] Attamimi Muhammad，中村友昭，長井隆行，“生活支援ロボットの遠隔操作からの自律化”，計測自動制御学会 システム・情報部門 学術講演会，pp.157–160，2013.

[12] Muhammad Fadlil，Attamimi Muhammad，長井隆行，中村友昭，船越孝太郎，“多層マルチモーダルLDAによる動きと物体の統合的概念の形成と語

意獲得”，計測自動制御学会 システム・情報部門 学術講演会，pp.161–165, 2013.

[13] 岩崎安希子，下斗米貴之，鴫原宏明，藤岡直幹，安東裕司，日永田智絵，アッタミミ ムハンマド，長井隆行，大森隆司，“生体指標によるロボット-子供遊び戦略の妥当性の評価”，HAI シンポジウム 2013，P4，pp.55–58，2013.

[14] 岩崎安希子，下斗米貴之，阿部香澄，鴫原宏明，安東裕司，日永田智絵，アッタミミ ムハンマド，長井隆行，大森隆司，“ロボット-子供遊び戦略と生体指標による評価”，第 9 回 日本感性工学会，2014.

[15] Attamimi Muhammad，長井隆行，岩橋直人，奥乃博，“音声命令と行動予測に基づく文脈を考慮したロボットの行動決定”，第 14 回 計測自動制御学会 システムインテグレーション部門講演会 SI2013，3K3-6，2013.

[16] アッタミミ ムハンマド，中村友昭，長井隆行，“移動ロボット知覚制御用 RTC 群の開発と学生実験での利用”，第 14 回計測自動制御学会システムインテグレーション部門講演会，1B4-5，2013.

[17] ブイターントゥン，ムハンマド アッタミミ，中村友昭，長井隆行，稲邑哲也，“曖昧性や身体的制約をベイジアンネットワークで統合するコミュニカティブモバイルマニピュレーション”，第 14 回計測自動制御学会システムインテグレーション部門講演会，1K3-2，2013.

[18] 日永田智絵，アッタミミ ムハンマド，長井隆行，下斗米貴之，大森隆司，“人とロボットのフィジカルコミュニケーション：手をつないで一緒に散歩するロボットの実現”，第 14 回 計測自動制御学会 システムインテグレーション部門講演会 SI2013，1J2-2，2013.

[19] 鴫原宏明，藤岡直幹，安東裕司，日永田智絵，Attamimi Muhammad，長井隆行，岩崎安希子，下斗米貴之，大森隆司，“サービスロボットのための遠隔操作システムの開発”，第 14 回 計測自動制御学会 システムインテグレーション部門講演会 SI2013，3A4-6，2013.

[20] Muhammad Fadlil，Attamimi Muhammad，阿部香澄，中村友昭，長井隆行，“多層マルチモーダル LDA を用いた多様な概念の統合と語意の獲得”，人工知能学会全国大会，1I3-3，2014.

[21] Attamimi Muhammad，中村友昭，長井隆行，岩橋直人，奥乃博，“ユーザ行動の予測と命令解釈の統合に基づくロボットの行動決定手法”，人工知能学会全国大会，1I4-OS-09a-1，2014.

[22] 西原成，中村友昭，アッタミミ ムハンマド，長井隆行，“物体概念と言語モデルのオンライン相互学習”，日本ロボット学会学術講演会，2I1-03，2014.

[23] Attamimi Muhammad，中村友昭，長井隆行，持橋大地，小林一郎，麻生英樹，“獲得した階層的概念・語意・文法に基づく文生成”，日本ロボット学会学術講演会，3I1-04，2014.

[24] 安東裕司，アッタミミ ムハンマド，中村友昭，長井隆行，持橋大地，小林一郎，麻生英樹，“日常生活言語化のためのデータ取得システム”，第 14 回計測自動制御学会システムインテグレーション部門講演会，3J4-5，2014.

[25] 山口拓郎，アッタミミ ムハンマド，中村友昭，長井隆行，池原雅章，“マルチモーダル LDA を用いたロボットによる感情語彙の獲得”，第 14 回計測自動制御学会システムインテグレーション部門講演会，1M4-5，2014.

[26] 鴫原宏明，アッタミミ ムハンマド，阿部香澄，長井隆行，大森隆司，岡夏樹，“他者モデルを有するエージェントインタラクションにおける二者関係のモデル化”，第 14 回計測自動制御学会システムインテグレーション部門講演会，2G2-2，2014.

[27] 片上祐介，阿部香澄，アッタミミ ムハンマド，長井隆行，“人とロボットの対話における正直シグナルの利用”，第 14 回計測自動制御学会システムインテグレーション部門講演会，2G2-4，2014.

受賞歴

個人受賞

[1] ロボカップ研究賞 2011：Muhammad Attamimi, Akira Mizutani, Tomoaki Nakamura, Komei Sugiura, Takayuki Nagai, Naoto Iwahashi, Hiroyuki Okada, and Takashi Omori，“Learning Novel Objects Using Out-of-Vocabulary Word Segmentation and Object Extraction for Home Assistant Robots”, ICRA 2010, pp.745–750, 2010.

[2] SI2013 優秀講演賞: Attamimi Muhammad，長井隆行，岩橋直人，奥乃博，“音声命令と行動予測に基づく文脈を考慮したロボットの行動決定”，第 14 回計測自動制御学会 システムインテグレーション部門講演会 SI2013，3K3-6，2013.

[3] SI2013 優秀講演賞: Bui Thanh Tung，Attamimi Muhammad，中村友昭，長井隆行，稲邑哲也，“曖昧性や身体的制約をベイジアンネットワークで統合するコミュニカティブモバイルマニピュレーション”，計測自動制御学会システムインテグレーション部門講演会，1K3-2，2013.

[4] SI2013 RTC 再利用賞: Attamimi Muhammad，中村友昭，長井隆行，“移動ロボット知覚制御用 RTC 群の開発と学生実験での利用”，計測自動制御学会システムインテグレーション部門講演会，1B4-5，2013.

[5] SI2013 ヴィストンロボットショップ賞: Attamimi Muhammad，中村友昭，長井隆行，“移動ロボット知覚制御用 RTC 群の開発と学生実験での利用”，計測自動制御学会システムインテグレーション部門講演会，1B4-5，2013.

団体受賞

[1] ロボカップジャパンオープン 2009 大阪@ホームリーグ 優勝，May. 2009.

[2] ロボカップ人工知能学会賞，May. 2009.

[3] ロボカップ世界大会 2009 グラーツ@ホームリーグ 準優勝, Jul. 2009.

[4] ロボカップジャパンオープン 2010 大阪@ホームリーグ 優勝，May. 2010.

[5] ロボカップロボット学会賞，May. 2010.

[6] ロボカップ世界大会 2010 シンガポール@ホームリーグ 優勝, Jun. 2010.

[7] ロボカップジャパンオープン 2011 大阪@ホームリーグ 優勝，May. 2011.

[8] ロボカップ人工知能学会賞，May. 2011.

[9] ロボカップジャパンオープン 2012 大阪@ホームリーグ 優勝，May. 2012.

[10] ロボカップ世界大会 2012 メキシコ@ホームリーグ 準優勝, Jun. 2012.

[11] ロボカップジャパンオープン 2013@ホームリーグ 3 位，May. 2013.

[12] ロボカップジャパンオープン 2014@ホームリーグ 優勝，May. 2014.

Book Chapters

[1] Muhammad Attamimi, Tomoaki Nakamura, Komei Sugiura, Takayuki Nagai, Naoto Iwahashi, “Learning Novel Objects for Domestic Service Robots,” The Future of Humanoid Robots: Research and Applications, ISBN 978-953-307-951-6 published by InTech.

謝辞

まず，的確な助言や多大なるご指導を頂いた指導教官である長井隆行教授に心より深く感謝致します．研究室に配属されたとき，研究の仕方は勿論のことプログラミングなど研究に必要な知識を丁寧に教えて下さいました．また，研究に行き詰まって何をしたら良いか迷ったとき，幅広い知識で解決策を一緒に考えて下さいました．さらに，自分の視野や研究に関する経験を広げるために，国内学会だけではなく憧れの国際学会に行くための機会を与えて頂きました．また博士論文の執筆に当たり，何度も相談に乗って頂くだけではなく，日本語のチェックまで親切にして頂きました．これ以外にも，ロボカップやロボットの展示会など数々の活動の機会を与えて頂き，貴重な経験をさせて頂きました．本当にありがとうございました．

また，ご多忙の中，学位論文の審査委員をして下さった，金子正秀教授，田中一男教授，横井浩史教授，内田雅文准教授に心より感謝致します．先生方の貴重なコメントやアドバイスを頂くことで，より良い論文を書くことができました．

そして，研究室に配属されたとき，研究の進め方や悩みについて親身になって相談に乗って下さった中村友昭先生に，この場をお借りして心よりお礼を申し上げます．

また，これまで研究室で共にした長井研究室のメンバーに感謝致します．私に様々なことを教えて下さり，相談に乗って頂いた水谷了さん，同期の伊東さん，板谷さん，共に研究するだけではなく，自分の悩みなどを聞いて下さった荒木さん，ファドリルさん，名前を挙げればきりがありませんが，研究室のメンバーのお蔭で有意義で楽しい時間を過ごすことができました．

最後に，これまでに素晴らしい学生生活を送らせて頂き，様々なところで私を日々支えて下さった家族に心より感謝の気持ちを申し上げます．

著者略歴

ATTAMIMI MUHAMMAD　（あったみみ　むはんまど）

1985 年 3 月 27 日	インドネシアに生まれる
2005 年 4 月	国立東京工業高等専門学校
	電子工学科編入学
2008 年 3 月	国立東京工業高等専門学校
	電子工学科卒業
2008 年 4 月	電気通信大学電気通信学部
	電子工学科編入学
2010 年 3 月	電気通信大学電気通信学部
	電子工学科卒業
2010 年 4 月	電気通信大学大学院情報理工学研究科知能機械工学専攻
	博士前期課程入学
2012 年 3 月	電気通信大学大学院情報理工学研究科知能機械工学専攻
	博士前期課程修了
2012 年 4 月	電気通信大学大学院情報理工学研究科知能機械工学専攻
	博士後期課程入学
2012 年 4 月	日本学術振興会特別研究員 DC1
2015 年 3 月	電気通信大学大学院情報理工学研究科知能機械工学専攻
	博士後期課程修了予定

電子情報通信学会，ロボット学会，人工知能学会学生会員